Gene regulation in microglia and genetic risk for complex brain disorders

A thesis submitted for the degree of Doctor of Philosophy
at
Cardiff University
School of Medicine

Darren Cameron
2019

Statements and Declarations

Statement 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Signed _____ Date _____

Statement 2

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is it being submitted concurrently for any other degree or award (outside of any formal collaboration agreement between the University and a partner organisation)

Signed _____ Date _____

Statement 3

I hereby give consent for my thesis, if accepted, to be available in the University's Open Access repository (or, where approved, to be available in the University's library and for inter-library loan), and for the title and summary to be made available to outside organisations, subject to the expiry of a University-approved bar on access if applicable.

Signed _____ Date _____

Declaration

This thesis is the result of my own independent work, except where otherwise stated, and the views expressed are my own. Other sources are acknowledged by explicit references. The thesis has not been edited by a third party beyond what is permitted by Cardiff University's Use of Third Party Editors by Research Degree Students Procedure.

Signed _____ Date _____

Word Count: 34,385

Acknowledgments

I would like to thank Nick, Matt, Manuela, Heath, Chris, Gareth, Derek, Katherine, Jo, Kate, Cath, Ann, Andrew, Ric, Antonio and the core team. In addition, I would like to thank the Medical Research Council for funding this work.

ACRONYMS

ADHD:	Attention-deficit hyperactivity disorder
ASD:	Autism spectrum disorder
ATAC-seq:	Assay of transposase-accessible chromatin with sequencing
BPD:	Bipolar disorder
ChIP-seq:	Chromatin immunoprecipitation with sequencing
CLOZUK:	Genotype sample of patients with treatment resistant schizophrenia who have be prescribed clozapine
C4AL:	Complement component C4 long form
CNS:	Central nervous system
CNV:	Copy number variant
DMSO:	Dimethyl sulfoxide
DNA:	Deoxyribonucleic Acid
EADI:	European Alzheimer's disease initiative
EDTA:	Ethylenediaminetetraacetic acid
ENCODE:	Encyclopaedia of DNA elements project
EMSA:	Electrophoresis mobility shift assay
eQTL:	Expression quantitative trait loci
ETS:	E twenty-six
FACS:	Fluorescence-activated cell sorting
FSC-A:	Forward scatter area
FSC-H:	Forward scatter height
GARFIELD:	GWAS analysis of regulatory or functional information enrichment with LD correction
GERAD:	Gene and environmental risk in Alzheimer's disease consortium
G-MCF:	Granulocyte/macrophage colony stimulating factor
GWAS:	Genome-wide association study
HDBR:	Human developmental biology resource
HM:	Histone modification
HOMER:	Hypergeometric optimisation of motif enrichment
IGAP:	International genomics of Alzheimer's disease project
IGV:	Integrative genomics viewer
iPSC_MG:	Immune pluripotent stem cell derived microglia
iPSC_MΩpre:	Immune pluripotent stem cell derived macrophage precursor cells
LD:	Linkage Disequilibrium

LOAD:	Late onset Alzheimer's disease
M-CSF:	Macrophage colony stimulating factor
MDD:	Major depressive disorder
OCR:	Open chromatin region
PCA:	Principal component analysis
PCR:	Polymerase chain reaction
PCW:	Post conception weeks
PGC:	Psychiatric Genomics Consortium
PWM:	Position weight matrix
qPCR:	Quantitative polymerase chain reaction
REMC:	United States Institute of Health Roadmap epigenetics consortium
RNA:	Ribonucleic acid
SCZ:	Schizophrenia
SEB:	Substrate equilibrium buffer
sLDSC:	Stratified linkage disequilibrium score regression
SNP:	Single nucleotide polymorphism
SNV:	Single nucleotide variant
SPD:	Schizophrenia patient derived
SSC-A:	Side scatter area
SV40:	Simian virus 40 large T antigen immortalised
TBE:	Tris-borate-Ethylenediaminetetraacetic acid
TF:	Transcription factor
TSS:	Transcription start site
vlPFC:	Ventrolateral prefrontal cortex

Summary

In recent years, genome wide association studies (GWAS) have established that common genetic variation plays an important role in complex brain disorders, such as autism spectrum disorder, schizophrenia and Alzheimer's disease. However, as the vast majority of GWAS risk loci are located in poorly characterised non-coding regions, interpretation of GWAS data is difficult. As very few associations can be attributed to effects on protein-coding sequence, the majority of common risk loci for complex brain disorders are believed to operate through effects on gene regulation.

In order to elucidate genetic risk mechanisms for complex brain disorders, it is important to establish in which cell types these risk loci are operating. Microglia are the primary immune cell of the central nervous system and are important regulators of brain function throughout life. As such, there has been growing interest in the potential role of microglia in brain disorders. One means of interpreting the function of the non-coding genome in a cell of interest is to measure the cell's chromatin landscape. For example, the assay for transposase accessible chromatin with sequencing (ATAC-seq) maps regions of 'open' chromatin, which have the potential to regulate gene expression through binding of transcription factors.

In this thesis, I use ATAC-seq to map open chromatin in adult *ex vivo* microglia, 2nd trimester foetal microglia, and *in vitro* human cell models of microglia. I then integrate this with complex brain disorder GWAS data to investigate whether gene regulatory processes in human microglia contribute to genetic risk for complex brain disorders. I find evidence that risk variants for late onset Alzheimer's disease, autism spectrum disorder, bipolar disorder, major depressive disorder and schizophrenia are enriched within regulatory regions utilised by foetal and / or adult microglia, suggesting a primary role for this cell type in these conditions. Using the electrophoretic mobility shift assay, I further show that two single nucleotide polymorphisms associated with Alzheimer's disease, and within regions of open chromatin in adult microglia, alter binding of protein from microglial nuclei.

Table of Contents

1 General Introduction

1.1 Definition of brain disorders

The practical and theoretical core of this thesis centres on leveraging functional genomic data derived from a particular brain cell to interpret genetic data derived from patients with disorders of the brain. These disorders include schizophrenia, autism spectrum disorder, attention deficit hyperactivity disorder, late onset Alzheimer's disease, bipolar disorder and major depressive disorder. For the purposes of this thesis, the terminology I will use to collectively describe these conditions is 'complex brain disorders', or simply 'brain disorders'.

At certain times during the text I use additional terms, such as psychiatric, neurodevelopmental or neurodegenerative, as a means to sub-characterise brain disorders; however, I acknowledge that these terms can be ambiguous and often mean different things in different settings. For example, in the latest release of the Diagnostic and Statistical Manual of Mental Disorders (DSM), DSM-V, a neurodevelopmental disorder is defined as a disorder where psychopathological symptoms have a childhood or adolescent age of onset (1). As such, autism spectrum disorder and attention deficit hyperactivity disorder are characterised as neurodevelopmental disorders, but, schizophrenia is not on the basis that psychotic symptoms for schizophrenia (usually) first manifest in early adulthood (schizophrenia is classified as a psychotic disorder in DSM-V; 1). By contrast, in the scientific literature, schizophrenia is often cited as a neurodevelopmental disorder, due a variety of epidemiological, neurobiological and now genetic data suggesting a prenatal component to this condition (2–5). In this thesis, I will refer to schizophrenia, bipolar disorder and major depressive disorder as 'psychiatric disorders', autism spectrum disorder and attention deficit hyperactivity disorder as 'neurodevelopmental disorders' and Alzheimer's disease as a 'neurodegenerative disorder'.

In the absence of robust biomarkers for these conditions, classification continues to be based primarily on a descriptive psychological examination (6). The core diagnostic features for each of the brain disorders I consider in this thesis are as follows.

Autism spectrum disorder (ASD) is a developmental disorder with a childhood onset and is characterised by deficits in social communication and behaviour (7). Prevalence of ASD is estimated at ~1-1.5% worldwide and affects children and adults (8,9). In children with ASD, diagnostic features can include delayed speech development, the use of simple repetitive phrases rather than structured sentences when verbalising or being unresponsive when spoken to or asked to perform a task. Non-verbal communication can also be diminished such that individuals make little or no eye contact, and do not gesticulate normally, when interacting with others. Children with ASD also have difficulty understanding and forging relationships and prefer to play alone. Regarding behavioural abnormalities, activities are usually stereotyped and highly repetitive, such as hand flapping or rocking to-and-fro continuously. Play is also unimaginative such that toys or objects are repetitively organised by colour or size. ASD patients also demonstrate an intolerance to any alteration in their normal routine and can exhibit extreme distress when normal rituals are deviated from (10).

Attention deficit hyperactivity disorder (ADHD) is a neurodevelopmental disorder that affects 5% of children and adolescents worldwide (11) and is characterised by inattentiveness and hyperactive or impulsive behaviour. Suffers may be unable to focus on specific tasks for extended periods and they constantly lose interest and make careless mistakes. Persistent fidgeting, excessive talking and movement, and acting without thinking are also common features. Symptoms often result in social and disciplinary problems causing isolation from their peer group and educational underachievement (12). For an ADHD diagnosis to be given, symptoms must present before the age of 12 (and be continuous for at least 6 months), they must occur in at least two environmental settings (i.e. at school and at home) and considerably impact social and educational progress. Although symptoms can partially or fully recede in adulthood it is estimated that 2.5% of adults remain symptomatic (13). Before adulthood, ADHD is 4-times more likely to affect males, but this sex-bias normalises after adolescence (11).

Bipolar disorder (BPD) is an episodic illness characterised by polarised alterations of mood, energy level and behaviour which manifest as periods of mania or depression. There are two major classifications of BPD: Bipolar I which involves at least one episode of mania, and BPD II which is characterised by at least one depressive episode and one hypomanic episode (i.e. a subthreshold period of mania; 14). During manic episodes, individuals with BPD are commonly hyperactive, have delusions of

grandeur and have a reduced need to sleep. In depressive episodes, patients become socially withdrawn, hypoactive and despondent. Psychosis is also a typical feature of BPD, most commonly during mania, but it can also occur during depression. Cognitive impairments are also sometimes seen, affecting executive function, attention and memory (15). BPD has a mean age of onset of ~20 years (16), and affects >1% of people worldwide (17).

Late onset Alzheimer's Disease (LOAD) is a progressive neurodegenerative condition. It is the most common form of dementia, accounting for 62% of all dementia cases in the UK (18). Relentless cognitive decline is a major symptom of the disease and, as the disease progresses, the severity of decline correlates with synapse and neuronal loss that begins first in the hippocampus before spreading widely throughout the brain (19–21). Symptoms in the early stages of the disorder include mild deficits in attention, problem solving and short-term memory that do not affect an individual's independence. However, over time these symptoms worsen to such a degree that individuals lose the capacity to complete basic daily tasks, like washing and dressing. In the latter stages of the disease, physiological and motor deficits emerge, leaving sufferers incontinent and unable to walk or speak (22). On average, symptom progression lasts 7-10 years with death being the end result. The defining neuropathological features of LOAD are insoluble extracellular amyloid plaques and intracellular neurofibrillary tau tangles (23).

Major depressive disorder (MDD) is a neuropsychiatric disorder characterised by episodes of depressed mood that last longer than 2 weeks which are not associated with an underlying health condition or stressful life event. Individuals with MDD can exhibit a marked reduction in their capacity to take pleasure in life and commonly experience feelings of worthlessness and guilt. As such, individuals with MDD are 20-times more likely to die by suicide than the general population. Sleep and appetite disturbance, fatigue and social detachment are also typical symptoms (24). MDD is twice as likely to occur in females than in males (25) and affects 1 in every 6 adults making it the second highest contributor to disease burden worldwide (26,27). The median age of onset of MDD is 25; however, symptoms can emerge in mid-adolescence (26). MDD is associated with periods of sickness and remission with episodes of illness ranging, on average, between ~3-7.5 months.

Neuroticism is personality trait that describes the tendency of an individual to respond with negative emotion to daily experiences. Individuals high in neuroticism frequently

respond with disproportionate negativity to life challenges and are disposed to experience their environment (or situation) as being stressful or threatening. Feelings such as personal inadequacy or dissatisfaction with the world at large are commonly expressed (28). Neuroticism is of interest in neuropsychiatric research as it associates with conditions such as schizophrenia (29) and depression (30).

Schizophrenia (SCZ) is a neuropsychiatric disorder that affects ~1% of the population (31). As the symptoms of SCZ are wide-ranging, and present heterogeneously, they are divided into three broad categories; namely, positive, negative and cognitive. Positive symptoms include thoughts or motivations that are additional to normal behavioural repertoire, such as auditory or visual hallucinations, which may cause an individual to lose their grasp of reality (32). Paranoid and delusional modes of thought often accompany hallucinations such that individuals have an unrealistic outlook and become convinced they are being pursued, controlled or persecuted. In contrast, negative symptoms are associated with reductive patterns of behaviour such that individuals become withdrawn, experience anhedonia or lack motivation (32). Cognitive dysfunction is also a core feature of SCZ with deficits in reasoning, information processing, planning and working memory being common (33,34). Cognitive symptoms usually manifest first during adolescence; however, diagnosis often does not take place until early adulthood upon presentation of the first psychotic episode (35,36). Often people with schizophrenia have disorganised speech and behaviour such that they fail to interact properly in social situations or at work, which can lead to long-term unemployment and social isolation. People with schizophrenia have reduced life expectancy (37) and are more likely to commit suicide than the general population (38).

Despite a century of research, definitive characterisation of the molecular events that cause the majority of these disorders remains elusive. As all of these disorders severely impair the cognition, behaviour and mood of those affected, they place a significant burden on health and social care systems worldwide.

1.2 Genetics of complex brain disorders

For many years, scientists have been interested in understanding how genetic variation between individuals contributes to complex brain disorders. Like the majority of human traits, many clinical disorders are heritable to some degree. Considering

the heritability of complex brain disorders, evidence from twin studies, which compare phenotype concordance between monozygotic and dizygotic twins, shows that a significant component of risk for most brain disorders is driven by genetic factors, (see table 1.1; 39,40). Indeed, in the case of schizophrenia, genetic inheritance is considered the strongest risk factor for the disorder (41,42).

Table 1.1 Heritability estimates of complex brain disorders

Disorder	Heritability estimate
Attention-deficit hyperactivity disorder	0.75
Autism spectrum disorder	0.80
Bipolar disorder	0.75
Late onset Alzhiemer's disease	0.60-0.80
Major Depressive disorder	0.37
Schizophrenia	0.81

Adapted from table in (42) except late onset Alzhiemers disease source (40)

However, risk for complex brain disorders is not mediated by genes alone, but by a combination genetic and environmental factors. The theoretical rationale for studying the genetic component of these disorders is that genes, and the genome as a whole, are fixed and measurable entities, and far easier to quantify in an unbiased, objective, manner than an individual's environmental experiences. Moreover, an individual's genetic landscape provides the biological outline for the traits that will interact with the environment. It encodes the resilience and/or vulnerability an individual has for a particular disorder and, as such, makes it a logical place to look for clues regarding the underlying pathological mechanisms of these disorders (43).

1.2.1 Early molecular genetic studies of brain disorders

Early efforts to identify genetic risk variants for brain disorders relied principally upon genetic linkage and candidate gene association studies (44).

Linkage studies are family-based studies that attempt to measure whether a disorder, and the alleles that cause it, co-segregate with known genetic markers that are situated throughout human chromosomes. As such, when co-segregation occurs

consistently within a family pedigree, it implies that a particular genomic region is a candidate risk locus for that disorder. Notable successes of genetic linkage studies were the identification of APP, PSEN1 and PSEN2 as risk genes for early onset Alzheimer's disease, a rare, more aggressive, form of the disease that affects individuals younger than 65 years (45–47). However, while linkage studies have been successful in identifying rare variants conferring strong effects on risk for certain (neurodegenerative) disorders (45,46) they have had limited success in identifying common risk variants for brain disorders (48).

In candidate gene association studies, the frequency at which certain genetic variants occur is usually compared between individuals with a disorder ('cases') and unaffected controls, with higher frequency in cases being taken as evidence that the variant is associated with the disease (49). These studies typically investigate a limited number of genetic variants at a candidate gene locus in a small number of individuals (usually < 1000). Although candidate gene association analyses were thought to be better suited to identifying common variants conferring weak effects on disease risk than linkage analyses, they rarely produced results that could be replicated across studies (50). An exception to this was the identification of the APOE epsilon 4 allele as an important risk factor for late onset Alzheimer's disease (51). Prior to genome-wide analyses, candidate genes were selected in a biased fashion based on imperfect biological assumptions. With the advent of genomic technology, and significant collaboration between research groups to pool genetic samples, a coherent representation of the genetic architecture of brain disorders emerged that explained why the design of earlier genetic studies proved, for the most part, inadequate. It is now apparent that, for most complex brain disorders, genetic risk is conferred by many common alleles that individually confer a small effect on risk as well as rarer alleles of potentially stronger effect. These findings will be discussed in the next sections.

1.2.2 Rare genetic variation

Rare genetic variation, typically defined as alleles with population frequencies less than 1%, can arise through small insertions or deletions of DNA sequence (indels), large structural alterations (copy number variants) or mutations at the base pair level (single nucleotide variants). Compared to common alleles, rare mutations are

considered to have arisen in recent ancestry (52). Disease-relevant rare mutations are the most difficult form of variation to detect due to extremely low allelic frequencies in the population and the large sample sizes required to detect them in multiple individuals. However, in recent years, exome, and whole genome sequencing technologies have made it possible to identify rare variants disrupting individual genes that occur more often in individuals with brain disorders than in unaffected controls.

Copy number variants (CNVs) are typically defined as deletions or duplications of at least 1kb. Several rare CNVs have been linked to increased risk for complex brain disorders (53). To date, 11 CNVs have been robustly linked to schizophrenia risk including six deletions and five duplications with odds ratios from 2 to >50 (54,55). Of these risk loci, CNVs on chromosomes 1q21.1, 13q13.3 and at the *NRXN1* locus have also been shown to increase ASD risk (56–58), suggesting shared pathways in certain psychiatric disorders. An increased rate of rare CNVs has also been observed in ADHD (59). However, CNVs appear to be less important in the aetiology of bipolar disorder (60) and major depressive disorder (61). Similarly, no CNVs have been robustly associated with LOAD, although duplication of the amyloid precursor protein has been linked with the early onset form of the disease (62).

Rare single nucleotide variants (SNVs) are the most difficult form of genetic variant to identify. It is estimated that, on average, ~74 *de novo* SNVs occur each generation (63). *De novo* SNVs have the potential to be more deleterious than inherited variants as they have undergone less selection pressure (64). Trio-based studies, which compare the exomes of affected cases and their parents, are commonly used to detect rare SNVs that arise between generations, and recent reports have associated *de novo* SNVs with complex brain disorders such as schizophrenia and autism spectrum disorder (65–67). Indeed, several rare LOAD-associated SNVs have been identified in the gene encoding the myeloid cell surface receptor TREM2 that increase risk 2-4 fold (68–70).

1.2.3 Genome-wide association studies

With the development of microarray-based genotyping technology in the early 2000s, together with improved understanding of haplotype structure and frequencies in

human populations (71,72), it became possible to perform genome-wide association studies of complex traits. These typically involve genotyping thousands of single nucleotide polymorphisms (SNPs) across the genome in large numbers of cases and controls or in a population that varies on a quantitative trait (e.g. plasma lipid levels). The advantage of this approach is that it provides a global screen for common variant association with a trait, without knowledge of the functional variants or bias towards prior hypotheses of underlying biology. However, the multiple testing burden is such that only associations where $P < 5 X10^{-8}$ ($P < 0.05$ Bonferroni-corrected for 1 million tests) are considered 'genome-wide significant'. In order to detect association of small effect alleles at such low P-values, large sample sizes are required. For brain disorders, like for many other complex traits, these have been achieved through the formation of large international research consortia.

1.2.4 Genome wide association studies of brain disorders

In 2009 the Genetic and Environmental Risk in Alzheimer's Disease Consortium (GERAD) and the European Alzheimer's Disease Initiative (EADI) simultaneously published two GWAS, implicating novel genetic loci for late onset Alzheimer's disease (LOAD) in *PICALM*, *CLU* and *CR1* (73,74). Despite the fact that both studies were carried out on independent samples, they both implicated loci containing the *CLU* and *APOE* genes, suggesting that the GWAS methodology could provide robust genetic associations. Until that point, alleles at *APOE* were the only common variants robustly associated with LOAD due to unusually strong effects of the common e4 allele (75). In subsequent years, several independent consortia published their own GWAS, identifying 6 additional LOAD genome-wide significant risk loci (76–78). In order to increase the statistical power to detect common variants with smaller risk effects, these consortia pooled resources to launch the International Genomics of Alzheimer's Project (IGAP). The IGAP's most comprehensive GWAS meta-analysis to date, using data from a sample of 21,982 cases and 41,944 controls, reported genome-wide significant associations between LOAD and 25 loci (79). Promisingly, 20 loci of these loci had been identified in a smaller meta-analysis carried out in 2013 (80). Moreover, the 5 newly reported loci had fell just below the genome-wide significance threshold in the previous study, demonstrating the benefit of increasing GWAS sample size to capture more of the common variant signal.

GWAS have also identified high confidence common risk loci for schizophrenia and other psychiatric disorders. In 2014, the Psychiatric Genetic Consortium (PGC) published data from the colloquially termed 'PGC2 study' reporting genome-wide significant associations between schizophrenia and 108 independent loci (81). Candidate genes contained within these loci implicated genes involved in glutamate, dopamine and calcium channel signalling. More recently, Pardinas and colleagues combined PGC2 data with that from the 'CLOZUK' sample, identifying a total of 145 genome-wide significant associations with schizophrenia (82). In recent years the PGC has published GWAS data for other psychiatric and neurodevelopmental conditions. Although the sample sizes collected so far do not match the level of those generated for schizophrenia, robust risk loci have been reported, including 16 genome-wide significant risk loci in ADHD (83), 12 in ASD (83), 30 in BPD (84), and 44 in MDD (85).

1.2.5 Pathway analysis

As complex brain disorders are highly polygenic, interpreting how genetic perturbations contribute to illness is challenging. Pathway analysis is an analytical aggregation method that determines whether specific biological annotations, such as system, cellular or molecular networks, are enriched for genetic risk loci, thus implicating these networks in risk for brain disorders (86,87). This method has been used extensively in brain disorder genetics research. For example, networks involved in the regulation of the immune response, endocytosis, cholesterol transport and protein ubiquitination have been reported to be enriched for LOAD GWAS risk variants (88). Similar analyses have been carried out leveraging rare schizophrenia risk loci. Using a case-control gene set enrichment analysis approach, Pocklington and colleagues reported that CNVs were enriched in genes associated with excitatory and inhibitory neurotransmission in people with schizophrenia compared to controls and in an exome sequencing study on schizophrenia trios, Fromer and associates reported that rare de novo mutations were enriched in genes associated with glutamatergic post-synaptic density proteins (65).

1.2.6 Limitations of genome wide association studies

Although GWASs have had undoubted success in identifying genetic loci that impact risk in brain disorders, these successes are tempered by several limitations that make aspects of GWAS data interpretation difficult.

For example, a limitation in terms of the interpretation of GWAS data is that assigning causality to any particular variant is problematic. This is due to the fact that GWASs are based on a heuristic methodology designed to measure genome-wide common variation by genotyping a relatively small number of SNPs. To understand GWAS methodology and why it is not possible to identify causal SNPs using GWAS data alone, it is necessary to understand two key evolutionary concepts. First, is a phenomenon called linkage disequilibrium (LD) which describes the tendency of SNPs in close spatial proximity to be inherited non-randomly. SNPs that are inherited together are described as being in LD and the genome is segregated into blocks of SNPs that are in LD. Second is that allelic variation within LD blocks in individuals from the same ethnic background (i.e. European) occurs at only a few critical SNPs; most alleles are identical. This makes it possible to 'map' common variation genome-wide by genotyping only the critical SNPs (referred to as index, or 'tagging', SNPs) in each LD block. Moreover, as only a limited number of haplotypes (i.e. a unique sets of alleles) exist at these critical SNPs within each population, the remaining non-index SNPs can be imputed post-hoc using the human reference genome. These concepts are fundamental in GWAS design as they make it possible to measure common variation in an individual's genome quickly and cheaply. The drawback is that as each LD block can contain >1000 SNPs, all of which could feasibly contribute to disease risk, it is not possible to conclusively distinguish functional risk-causing SNPs, from risk-associated SNPs (89). This means additional, epigenomic and/or fine-mapping analyses are required to attempt to identify SNPs that increase risk (89).

A second challenge when interpreting GWAS data is that the large majority of risk loci fall in non-protein coding regions of the genome (90). This means that it is not immediately obvious through which gene(s) or cell type(s) risk variants are operating. However, it does strongly implicate the regulation of gene expression in the aetiology of brain disorders (see section 1.3.6). As such, functional annotation of GWAS data with genome-wide epigenetic data is needed to ascertain the regulatory functions impacted by these loci (91,92).

1.3 The regulation of gene expression

Despite the fact that virtually all cell types in the human body contain the same genetic information, which is encoded in their DNA sequence, cells are morphologically and phenotypically diverse. Phenotypic diversity in cells, with the same underlying genetic template, is made possible by altering the regulatory architecture surrounding DNA. Complex networks of proteins, DNA elements and non-coding RNAs interact to determine the genes that are expressed in a particular cell type at a particular time. This control of gene expression is vital for the maintenance of cellular identity, but it also allows cells the flexibility to activate different groups of genes in response to changing environmental demands or at critical periods during their development.

1.3.1 Mapping the human epigenome

After the successes of Human Genome Project, it became apparent that only ~1% of the human genome encodes proteins. As such, there was a requirement to determine what the remaining 99% of the genome did. Whilst it was hypothesised that a crucial function of the non-coding part of the genome was to regulate gene expression, until that point, there had been no attempt to empirically confirm this on a genome-wide scale. The Encyclopaedia of DNA Elements project (ENCODE) project was created in 2003 to take on this challenge and its remit was to characterise, map and create a comprehensive repository of all the epigenetic features in the human genome (93,94). To achieve this, the consortium committed to developing new high throughout methods, technology and analysis strategies to streamline this process. Since its inception, the ENCODE project has produced seminal data highlighting the importance of chromatin organisation in the control of gene expression and, specifically, how discrete chromatin states (and molecular signatures) can functionally partition the genome in a cell-specific manner. Moreover, it was shown that by integrating cell-specific epigenomic data with GWAS data, that GWAS risk loci segregated in a non-random manner in DNA elements, or chromatin annotations, in cell types that are relevant to the GWAS phenotype, i.e. SNPs associated with Crohn's disease, an auto-immune disorder, were enriched in open chromatin regions of immune cells (90).

In recent years, many groups have produced work exploring non-coding epigenomic phenomena. The work of ENCODE and similar groups, such as the US National Institute of Health Roadmap Epigenomics Mapping Consortium (REMC; 97,98), provided the foundation for this work which has advanced our understanding of the regulatory networks involved in the control of gene expression and how they contribute to disease. This section will describe the chromatin and DNA features that are involved in epigenomic control of gene expression and how they influence complex brain disorders.

1.3.2 Chromatin and chromatin accessibility

Chromatin is a complex of DNA, RNA and protein that is located in the nucleus of all human cell types, and it is the main constituent of chromosomes. The core structural element of chromatin is the nucleosome, and each nucleosome consists of a 147 base pair section of DNA sequence that is coiled around 8 histone proteins. Nucleosomes are connected to one another via a short (~60bp) sections of DNA, called linker DNA, to form a repeating structure that resembles beads on a string. Each histone protein octamer is composed of two copies of the H2A, H2B, H3 and H4 histones, and these are bound on their external surface by a H1 linker histone (97,98). During cellular mitosis, chromatin tightly condenses such that it can be visualised microscopically. In order to regulate cellular gene expression, chromatin dynamically alters its physical state. Euchromatin, is relaxed, uncoiled, chromatin where histones are either well-spaced, or displaced entirely, exposing DNA to molecules (e.g. transcription factors; TFs) that regulate gene expression. By contrast, heterochromatin is tightly compacted, and histones are bunched together making DNA relatively inaccessible. Chromatin accessibility is a measure of the functional state of chromatin and is defined by the degree to which DNA is exposed. In any given cell type, accessible DNA constitutes ~2-3% of the entire genome (99). In order for a gene regulatory event to take place chromatin must be uncoiled and in the euchromatic state to allow regulatory molecules to interact with DNA. When chromatin is heterochromatic most regulatory molecules cannot get access to DNA due to phenomena such as steric hindrance (100,101).

The assay for transposase-accessible chromatin using sequencing (ATAC-seq) is a popular method used to measure genome-wide chromatin accessibility (see figure

1.1; 102). This method takes advantage of the fact that euchromatin is sensitive to enzymatic cleavage. In the ATAC-seq assay, nuclei are extracted from a cell of interest and incubated with a hyperactive Tn5 transpose enzyme, which selectively excises exposed DNA from all the open chromatin regions of the genome. Crucially, the enzyme has been modified to carry adapter sequences such that, whilst excising the DNA, adapters are simultaneously inserted at the ends of the DNA fragments. These adapters act as barcodes for the DNA fragments so they can be recognised after pooled sequencing. Due to the high efficiency of the Tn5 enzyme, ATAC-seq has largely superseded alternative methods for measuring open chromatin regions, such as DNAse-seq (101). Once the DNA has been excised and barcoded, the DNA fragments are sequenced and the reads produced are aligned to a reference genome. The final read out of an ATAC-seq assay is a genome-wide set of 'peaks', representing sites of open chromatin that are potentially involved in the regulation of gene expression in the cell type of interest at the time the nuclei were extracted.

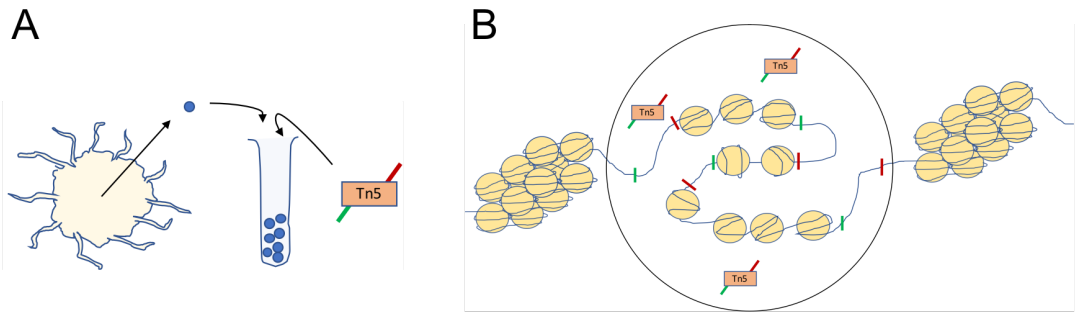

Figure 1.1. Diagrammatic representation of the ATAC-seq assay. (A) Nuceli are extracted from the cells of interest (microglia shown) and incubated with a hyperactive tn5 enzyme (orange) that has been modified to carry adapter sequences (shown in green and red). (B) When the modified tn5 enzyme enters microglia nuclei it recognises regions of the genome where chromatin is open and DNA is exposed. It then selectively excises exposed regions of DNA and simultaneously adds the sequencing adapters to the ends of the excised DNA. DNA fragments are then sequenced, and sequenced reads are aligned to a reference genome (hg19). This results in read pile ups at specific genomic locations called 'peaks'. Each peak represents a genomic region where chromatin is open. Globally, these peaks represent the regions in the genome that have the potential to regulate gene expression in microglia.

As described in the following sections in this chapter, open chromatin regions may be located either near the transcription start sites of genes or in non-coding regions located far from genes. This provides information on the genes that are likely to be active, or poised in preparation to be activated, within a cell type and the regulatory elements that have the potential to influence gene expression in these cells. It is also possible to identify specific DNA recognition sequences that gene regulatory proteins bind to within these loci. This provides a picture of the DNA-protein interactions that

regulate gene expression in the cell of interest. Furthermore, when integrating chromatin accessibility data with other functional genomic data it is possible to make functional predictions of the pathologically relevant effects that genetic variation may have in particular tissues or cell types.

1.3.3 Regulatory elements

Regulatory elements are non-coding sections of DNA sequence that act as binding platforms for DNA binding proteins, and they are involved in the initiation or modulation of gene expression (see figure 1.1). Cell-specific gene expression is possible through exposure of unique sets of regulatory elements to transcriptional proteins and co-factors. This ensures that cells have a unique set of activated genes that cater to their environmental needs, and distinct regulatory control over the rate of expression of those genes. Regulatory elements can be located proximally or distally to the genes that they influence, and several types have been described in the literature including promoters, enhancers, silencers and insulators. However, I will only discuss promoters and enhancers in this section.

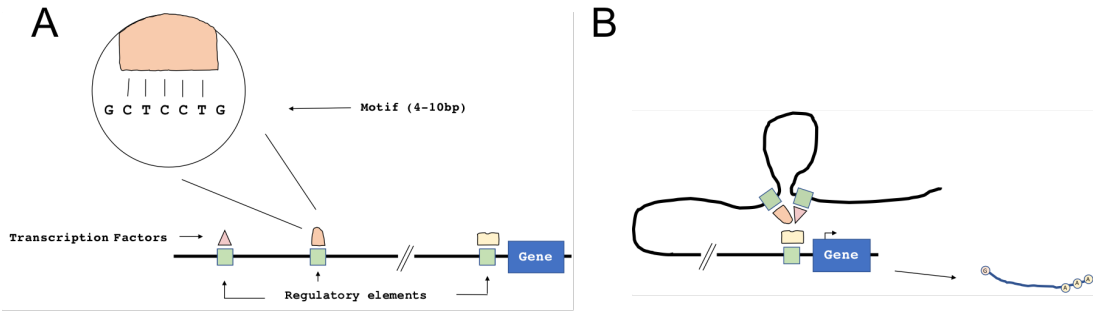

Figure 1.2. Diagrammatic representation of the relationship between regulatory elements and transcription factors. (A) Regulatory elements are non-coding sections of DNA that act act binding platforms for proteins called transcription factors (TFs) that are involved in the regulation of gene expression. DNA binding TFs bind to DNA in a sequence specific manner and recognise short sections of DNA called motifs. Promoters are regulatory elements that are situated proximal to genes whereas enhancers are located distally. (B) Enhancers and their associated TFs interact with promoters via chromatin looping to modulate expression of their target gene(s).

Promoters are a type of regulatory element that are located proximal to the genes they influence and are necessary for RNA synthesis (see figure 1.1). Whilst all genes have at least one promoter situated upstream of the 5' end of the gene, over 50% of genes have multiple, alternative promoters (103,104). The DNA sequence of a

promoter acts as a docking site for the transcription pre-initiation complex which, in turn, directs RNA polymerase II to the transcription start site of the gene to initiate mRNA production (105). Binding of RNA polymerase II is sufficient to drive low-level, basal transcription of RNA, so additional activator proteins that bind upstream of the core promoter, are required to modulate the rate of gene expression. Genes with alternative promoters can produce different protein isoforms by producing different mRNA transcripts thus contributing to phenotypic diversity (106,107). Moreover, alternative promoters from the same gene demonstrate tissue specificity (108,109) and have contrasting activity levels during development (110).

Unlike promoters, enhancers are located distally to the genes they regulate. Typically a few hundred base pairs in length, enhancers also act as binding platforms for transcription factors and, in general terms, increase the rate of expression of their target genes (111). Enhancers interact with transcription start sites of promoters through chromatin looping; a biophysical phenomenon where sections of genomic sequence form loops to bring distant regulatory elements into close spatial proximity with their target gene(s) (see figure 1.1; 112). Typically, enhancers are more difficult to identify than promoters as they lack the universal sequence characteristics, such as CpG islands (113) or shared protein binding elements (105), that help distinguish promoters across tissue types. It is also difficult to accurately predict the genes that enhancers target, particularly as a single enhancer can modulate the activity of multiple genes. Chromosome conformation capture (3C) - based assays have been developed to map these long-range regulatory interactions, although such datasets are not yet available for all cell types (and none currently exist for microglia).

Crucially, genetic changes that impact regulatory elements can contribute to variance in phenotypic traits and disease susceptibility between individuals. While only ~1% of the human genome encodes protein, up to 10% is evolutionarily conserved, suggesting that a significant portion of the non-coding genome is of functional importance (103).

1.3.4 Transcription factors

The term transcription factor (TF) is used to describe a protein that is involved in the regulation of gene expression. Transcription factors can influence gene expression

through direct interaction with DNA, or by binding indirectly to DNA as part of a larger protein complex. DNA binding transcription factors bind to DNA in a sequence specific manner, recognising short 4-12bp sections of DNA sequence called motifs (see figure 1.2; 114). A motif can be interpreted as the DNA footprint for a DNA binding transcription factor. Some transcription factors, such as those that form the transcription initiation complex, are generically expressed and have the same function across cell types, whilst other TFs are cell type specific. However, it is also possible for the same transcription factor to influence expression of a distinct set of genes in different cell types (116).

Cellular control of gene expression is critical in order to respond to internal and external stimuli and to balance two seemingly competing needs (A) the need to alter gene expression (e.g. at critical times during its developmental life cycle or to respond to environmental stimuli), which is particularly true for functionally dynamic cells such as microglia, and (B) the need to preserve a gene expression pattern, in order to maintain tissue homogeneity and cell identity (117).

One mechanism that is proposed to underpin these processes in cells is enhancer selection. The human genome is predicted to have hundreds of thousands of enhancer regions (118). The method by which cells preferentially select a specific group of enhancers in order to confer cell identity and/or to react to an environmental stimulus has been a topic of some debate. A consistent finding across studies is that combinations of transcription factors interact in a context-specific manner, spatially and/or temporally, to drive enhancer selection. For example, the pioneer factors PU.1 and C/EBP-α/-β interact in spatial context, to establish myeloid, and in particular macrophage, cell lineage (119). These pioneer factors, directly compete with histones for the DNA binding sites in heterochromatic regions (120). In macrophages, PU.1 binds specifically at enhancer sites that also contain C/EBP factor binding sites and these factors act in combination to induce histone depletion, increase chromatin accessibility and increase the deposition of active chromatin modifications, such as H3K4Me1, at macrophage enhancers (121–123). Once enhancers are exposed, they can be primed, or activated, by additional transcription factors to modulate transcription in their target genes. This illustrates the interplay that occurs between histones, DNA and transcription factors to drive cell-specificity and regulate gene expression; however, other factors such as biochemical modifications on the surface of chromatin also play a critical role in these processes.

1.3.5 Chromatin modifications

The surface of nucleosomes is decorated with chemical markers, referred to as chromatin modifications, that can disrupt the contacts between neighbouring nucleosomes to influence chromatin accessibility and recruit proteins which initiate key cellular processes such as DNA transcription, replication or repair. Chromatin modifications can be attached to histones or DNA and they demarcate the genome into distinct functional domains (124).

Histone modifications (HM) are post-translational modifications that are covalently bound to the amino acids of histone proteins. Structurally, histones are predominantly globular, but they also have a freely protruding N-terminus tail to which the vast majority of HMs are attached. HMs can act to dynamically to alter the structure and function of chromatin by direct interaction with DNA, or by recruitment of chromatin remodelling proteins, to influence gene expression. HMs can be mapped to functionally annotate the genome and to predict the nature and activation status of the regulatory elements that they are proximal to (125,126). Two histone modifications commonly used to annotate the human genome are methylation of histone 3 lysine 4 (H3K4Me1-3) and acetylation of histone 3 lysine 27 (H3K27ac). Histone methylation is a reversible process that is facilitated by histone methyltransferases, which attach methyl groups to histones, and demethylases, which detach methyl groups. The methylation status of H3K4 is used to demarcate proximal and distal regulatory elements (127). For example, H3K4Me3 is associated with active transcription at promoters and gene bodies (128,129), whereas H3K4Me1 correlates strongly with functional enhancers (130).

Chromatin immunoprecipitation with sequencing (ChIP-seq) is the assay used to detect histone modifications. In this assay, genomic chromatin is fragmented then incubated with antibodies that target a histone modification of interest. Chromatin fragments associated with that modification are then immunoprecipitated and the DNA in each fragment is sequenced (131). Similar to the data produced in the ATAC-seq assay, reads are aligned with a reference genome and regions where reads pile up (‘peaks’) indicate the genomic loci associated with that particular histone modification.

As well as post-translational modifications to histones, the DNA itself can be adorned with epigenetic marks. DNA methylation is an epigenetic modification that is

associated with the repression of transcription in mammals. This process is heritable and is facilitated by a family of DNA methytransferases which catalyse the transfer of a methyl group to the 5th carbon of the cytosine ring in DNA. In regions where regulatory elements are silenced, DNA is often directly methylated, and the chromatin is closed and compacted, so transcription factors can't access regulatory elements (132,133).

One of the main benefits of chromatin-based assays such as ATAC-seq and ChIP-seq is that they make it possible to functionally annotate the genome. This provides information on which regulatory elements (and which genes) are likely to be active within tissues or cell types, but importantly from the perspective of making sense of GWAS data, it provides information on the function of non-coding regions of the genome (134).

1.3.6 Integration of genetic and epigenomic data

A primary goal for groups working in brain disorder research is to identify the cell-types and molecular mechanisms that are important in these disorders in order to identify potential targets for pharmacological intervention. One approach used to do this is to integrate GWAS data, which provides information on the loci associated with a particular brain disorder, with cell-specific epigenomic data which contains details on the regions that are active and inactive within a particular cell type, and what these regions do. Theoretically, this makes it possible to functionally annotate GWAS risk loci and potentially identify the cell-specific regulatory elements they may perturb.

An example of this approach is a study by Fullard and colleagues, who measured chromatin accessibility of neurons and non-neurons across 14 distinct brain regions, and using a statistical method called stratified linkage disequilibrium score regression (described in section 2.2.4), integrated this with GWAS data from 3 complex brain disorders (SZ, ADHD and LOAD). They reported that neuronal open chromatin regions in the neocortex and striatum were enriched for SZ-associated common variants, with the hippocampus, nucleus accumbens and the superior temporal cortex being the regions most enriched (135). Pinpointing the specific SZ risk loci that operate in neurons in this manner demonstrates the benefit of functional annotation of GWAS data to draw biologically meaningful conclusions regarding the cell-types

(and brain regions) important in complex brain disorders. Indeed, ascertaining the cell-types that brain disorder risk variants operate in will be an important step toward unravelling the molecular aetiologies that drive these disorders.

1.4 Microglia

Microglia are the primary immune cell of the central nervous system (CNS) and in human adults constitute ~0.5-16% of the total brain cell population, depending on region (136). Microglia were first definitively characterized by del Rio-Hortega who catalogued their migratory and phagocytic properties as well as their distinctive ramified morphology (137). However, it is unlikely del Rio Hortega could have foreseen the diverse and complex range of functions microglia satisfy during the human life cycle. In recent years, it has emerged that microglia contribute significantly to critical processes such as immune surveillance, neurodevelopment and CNS homeostasis (138). As such, microglia are capable of responding to a wide range of environmental and internal stimuli and can tailor their gene expression profile to present the phenotype best suited to the conditions that they encounter in their local environment. This requires epigenomic activation of specific enhancers and transcription factors (139–141). Moreover, as our understanding of microglial function has grown, there has been increasing interest in their role in the neuropathology of complex brain disorders (142).

1.4.1 Microglia ontogeny and turnover

For many years the developmental origin of microglia was unclear. Their emergence early on in brain development suggested they derived from an embryonic progenitor, whilst their phenotypic and antigenic similarity to myeloid cells found in the peripheral tissues implied that they are derived from circulating blood monocytes (143). However, recent lineage-tracing studies in transgenic mice have driven the current consensus view that microglia arise from a unique, and transient $RUNX1^{+}$/c-Kit^{+}/$CD45^{-}$ erythromyeloid progenitor population in the embryonic yolk sac (143,144). Once the circulatory system has been established, these cells begin to infiltrate the neuroepithelial tissues of the primitive brain and upregulate microglia lineage-specific transcription factors (Pu.1 and Irf8) before finally migrating, and proliferating, in CNS

parenchyma. Importantly, due to the formation of the blood brain barrier, this microglia progenitor population remains distinct from the main pre-natal haematopoietic reservoir in the foetal liver, which is the primary source of proliferating peripheral myeloid cells until the development of the bone marrow (144–146). Although fate mapping studies of a similar nature have still to be carried out in humans, immunohistochemical studies do hint at a similar human microglial developmental trajectory (147–149).

Once the blood brain barrier has been established, it is hypothesised that, in normal physiological conditions, microglia remain isolated from the peripheral myeloid cells that circulate in the blood (150,151). Evidence supporting this comes from studies in parabiotic mice which show that when the circulatory systems of genetically identical animals are combined, whilst peripheral myeloid cells derived from both animals were found in the vasculature, only host microglia were found in the brain of each mouse (144,150,152,153). As a consequence of this, and in order to maintain adequate brain coverage in homeostatic and immune activated conditions, microglia are long-lived, compared to peripheral myeloid cells (154), and self-renewing (150).

In certain experimental circumstances, bone-marrow derived myeloid cells can enter brain parenchyma. For example, when the myeloid cell lineage-determining transcription factor Pu.1, or the cell surface receptor Csf1r, are genetically depleted in mice, peripherally engrafted bone-marrow derived cells from donor animals can infiltrate the brain of recipient mice (155–157). Although these engrafted cells are morphologically similar and express similar cell surface markers, they are transcriptionally and functionally distinct from microglia, despite the fact that they inhabit the same environmental niche in the brain (156,157). Findings such as this has led to speculation that future exploitation of this engraftment mechanism could lead to the use of peripheral, bone-marrow derived cells to treat brain disorders (156,158,159). Indeed, brain engraftment of bone-marrow derived macrophages has been shown to alleviate the symptoms of autism and Rett syndrome in mouse models (160,161).

The impact that ontogeny has on the intrinsic phenotypic repertoire of microglia, and myeloid cells generally, remains an open question in the field, however it will be important to clarify whether or not microglia are uniquely vulnerable to brain disorder-associated insults and/or may be uniquely targeted for therapeutic benefit (162).

1.4.2 Microglia's immunological functional repertoire

Microglia have two primary functions in the central nervous system: to provide immunological defence and to maintain a relatively constant homeostatic balance. In the absence of an infectious or immune insult, microglia maintain the homeostatic equilibrium of the brain parenchyma. In these conditions, microglia have a ramified morphology and they constantly extend and retract their processes to survey, and 'sample', their local microenvironment for chemical or proteinaceous stimuli that could elicit a context-specific response (163–165). Similar to macrophages in the peripheral tissues, microglia act as the principal responders of the innate immune system by monitoring the parenchymal landscape for biochemical signals. These signals can be brain-derived, alerting microglia to cellular waste or apoptotic cells marked for phagocytic removal, or they can arise from external sources, such as infectious organisms, initiating an innate immune response. When exposed to a response-inducing stimulus, microglia retract their processes, adopt an amoeboid morphology, and alter their gene expression profile to provide the tools required to resolve the situation.

As the first line of defence for the innate immune system, microglia have an array of receptors on their cell surfaces through which to detect invading pathogens. For example, microglia carry a family of transmembrane molecules called toll like receptors which recognise a wide array of pathogen-specific ligands, such as lipoproteins, on the surface of bacteria, viral RNA, and chemokines released by other immune cells such as INF-γ (166). Once detected, the transcription factor nuclear factor-κB is activated which orchestrates the upregulation of pro-inflammatory genes to increase the production of pro-inflammatory cytokines (i.e. TNF-α, IL-1, IL-6, IL-12, IFN-$\alpha;\beta$), chemokines (monocyte chemotactic protein-1, CXCL9-10) and free radicals (nitric oxide) to alert, and recruit, neighbouring microglia (and T-helper cells) to the threat, and efficiently eliminate the pathological disturbance. Microglia's antigen presenting capacity is also significantly increased which alerts cells of the adaptive immune system, such as T-cells and B-cells to the threat. However, a drawback of this swift and aggressive response is that some of the chemicals produced by microglia as part of this response, such as nitric oxide (167,168), are neurotoxic so have the potential to collaterally damage surrounding neurons. Microglia have also been associated with a range of anti-inflammatory, activation states (M2a, M2b, M2c) whereby arginase is released, and the presence of specific interleukins can (A)

induce tissue repair via expression of neurotrophic factors, (B) promote the release of the neuromodulators (Il-4, Il-10 and TGF-β) to inhibit neighbouring cells from releasing pro-immune factors or (C) induce a phagocytotic profile to remove dead cells and cellular debris (169–172).

However, the nomenclature surrounding microglia activation states has been questioned recently. The terms M1 and M2 derive from studies in peripheral macrophages and are used to describe their polarised pro- and anti-inflammatory activation states. Although this terminology has been widely adopted in the microglia literature, it has been criticised for being too simplistic. Moreover, as much of the evidence to support this work was carried out in microglial cell cultures *in vitro*, which cannot model the complex interactions that occur *in vivo*, there is little evidence that these binary activation states occur *in vivo*. Indeed, microglia are often exposed to competing stimuli simultaneously *in vivo,* which leads to co-expression of pro- and anti-inflammatory factors (173–176). This has led to calls for new, unbiased, classifications for states of microglia activation based on transcriptomic and/or epigenomic profiles (177).

1.4.3 Microglia's brain architectural functional repertoire

In addition to their immunological and homeostatic roles, microglia are also involved in refining neuronal circuitry and modulating synaptic function throughout life. Indeed, as microglia emerge from the yolk sac at the same time as the birth of early neurons, microglia's role in neural refinement is thought to be particularly important in the developing brain where extensive rewiring of the nervous system takes place (178,179).

As the predominant phagocytic cell in the brain, microglia have been implicated in the removal of developing, and mature neurons, in regions of the brain where neurogenesis takes place. This can be a passive process whereby non-activated microglia simply engulf and clear the remains of apoptotic neurites (180), or microglia may drive neuronal apoptosis via the release of reactive oxygen species, nerve growth factor or tumour necrosis factor (181–183). Microglial control of neurogenesis is particularly important during CNS development where they control the rate of turnover of neural precursors. For example, activated microglia have been shown to

colonise the interface between ventricular and subventricular zones in the pre-natal developing cortex in human, macaque and mouse and, in macaque, microglia consume and phagocytose neural progenitor cells developing in these proliferative zones (184). Furthermore, pharmacological deactivation, or elimination, of microglia increased the number of neural progenitor cells in the proliferative zones, indicating that microglia have a critical role in restraining neuronal production. Microglial control of neurogenesis continues into adulthood where they have been shown to eliminate apoptotic neural progenitor cells in mice (185).

Similarly, microglia have a pivotal role in sculpting the connections between neurons in a process called synaptic pruning, in which neuronal activity and the complement system are also involved (186). Complement is a group of proteins that the innate immune system uses to tag apoptotic cells or foreign bodies for removal. These tags are recognized by microglia which, in turn, engulf and remove tagged elements from the microenvironment. For example, in the developing mouse visual system (P5-P10), the complement factors C1q and C3 are widely expressed and they co-localise at immature, or weakly signalling, retinal ganglion cell synapses to tag them for removal. Crucially, when retinogeniculate synaptic density was measured in C1q and C3 KO mice there was excess synapses in KO animals compared with controls. Excess cortical neuronal connectivity has also been reported in C1q knockouts (187), and microglia have been shown to actively engulf synapses in the mouse hippocampus (188). Moreover, microglia are the primary cell type in the brain that express the CR3 receptor and the C3-CR3 interaction is required for synaptic engulfment and elimination of C3 tagged synapses (189). Other factors have been implicated in microglial mediated synaptic pruning. For example, during neurodevelopment, fractalkine (CX_3CL1) is expressed by neurons (190) and it has a chemotactic influence on microglia and may act to attract microglia to weak synapses (191–193). Microglia are the only cells in the CNS to expresses the fractalkine receptor (194,195) and mice lacking the receptor have increased neuronal connections in the hippocampus compared to wild-type mice (188). Similarly, the cytokine IL-33, expressed by astrocytes, has been shown to increase microglial engulfment of developing synapses in mouse thalamus and spinal cord via activation of the microglial cell surface receptor 1L1RL1. *IL-33* KO mice have excessive excitatory synapses (196). Furthermore, when the microglial immune receptor triggering receptor expressed on myeloid cells 2 (*TREM2*) is genetically deleted, the CA1 region in the hippopcampi of mice has increased excitatory neurotransmission and synapse density compared to wild-type littermates (197).

As well as synaptic removal, it has been postulated that microglia are also involved in synapse formation and modulation. In young adult mice lacking microglia, learning-associated spine formation and behavioural task scores, have been reported to be significantly reduced compared to wild-type animals (198). Microglia also have a role in modulating the activity rate of synapses. For example, in zebrafish, microglia have been shown to preferentially interact with highly active neurons to attenuate their activity in a contact-dependent manner (199). It has been proposed that microglia identify neurons with high activity levels through the P2RY12 receptor that detects ATP released after neuronal activation (200).

Microglia are also reported to interact with oligodendrocytes in the brain and support myelinogenesis. For example, after a demyelination event, microglia clear fatty debris from the injury site and secrete factors to modify the surrounding extracellular matrix and recruit oligodendrocyte precursor cells which support myelin regeneration (201).

1.4.4 Microglia and brain disorders

As scientific interest has grown, and a fuller understanding of the range of physiological processes that microglia are involved in in the brain has emerged, research into the role that microglial dysregulation plays in the aetiology of complex brain disorders has also developed (202). Indeed, many studies have reported a link between microglial function and the pathophysiology of late onset Alzheimer's disease. For example, microglia are consistently shown to engulf, proliferate and accumulate around amyloid plaques in patients with the disease (203,204) and the proportion of activated microglia in the brain increases as the disease advances (23,205). The amyloid plaque burden in the brain is an established biomarker for the disorder and one proposal is that inefficient clearance of these plaques by microglia leads to plaque build-up, which, over time, triggers a chronic inflammatory response that leads to synaptic and neuronal degeneration (206). However, debate is ongoing around whether microglia are assuming a protective role in this process and merely being overwhelmed when attempting to clear the plaques or whether they have a direct role in neurodegeneration by releasing cytotoxic factors that damage surrounding neurites (207). Genetic evidence also strongly implicates microglia function in LOAD pathogenesis. For example, the majority of genes implicated by

LOAD GWAS, including *CD33*, *ABI3*, and *PLGC2*, are preferentially expressed by microglia, and pathway analyses implicate immune and microglia-specific gene networks with increased risk. Common variants in the microglia lineage determining transcription factor PU.1 have been associated with reduced risk for LOAD (208) whilst rare variants in *TREM2*, encoding an extracellular myeloid cell receptor are associated with increased risk. Moreover, a study leveraging gene expression data from post-mortem brain tissue in a LOAD case-control sample identified genes related to immune networks, and specifically, microglial networks, as being associated with LOAD pathophysiology (209).

Microglia have also been hypothesised to be important in the pathophysiology of psychiatric and neurodevelopmental disorders. For example, it has been postulated that excessive synaptic pruning occurring during adolescence or early adulthood increases schizophrenia risk, and recent work implicates microglia (and the complement system) in this process (210). For example, Sekar and associates reported that a proportion of the common variant SZ risk is driven by structural alleles of the complement component C4 gene such that a higher copy number of the C4A long form (C4AL) haplotype is associated with increased C4 expression and increased SZ risk (211). C4 is released by neurons and it activates another complement molecule in the brain, C3, which is secreted by microglia. When activated, C3 tags synapses for elimination and promotes phagocytic engulfment of synapses by microglia (186,189,212). In a recent induced pluripotent stem cell study, SZ patient-derived (SPD) microglia had increased uptake of synaptic material, and SPD neurons grown in co-culture had reduced spine density, when compared to cells derived from healthy controls (213). Increased C3 deposition on SPD neurons was also reported to correlate with a higher copy number of C4AL genotype. As low spine density is a common neuropathological feature reported in the post-mortem brain of SZ patients (214), this work has led to the proposal that increased C4 release in SZ patients with more copies of C4AL increases activation of C3 that, in turn, promotes pathologically increased synaptic engulfment by microglia. Intriguingly, clinical studies suggest that the antibiotic minocycline, which inhibits microglial activation (215), may alleviate the negative (and potentially positive) symptoms associated with schizophrenia (216,217). Minocycline has been shown to decrease microglial engulfment of synapses *in vitro* (213).

Synaptic dysfunction has also been linked to the pathophysiology of ASD (218,219). For example, de novo CNVs identified in individuals with ASD have been shown to

contain genes that converge on synaptic networks (220) and neuroimaging studies have reported white matter overgrowth in cortical areas associated with social communication (221). Given microglia's role in synaptic removal and/or maturation during neurodevelopment their role in ASD aetiology has been investigated. Post-mortem studies have shown that microglia are more ramified, have larger cell bodies and are more densely populated in the cortex of ASD patients compared to age-matched controls. Moreover, a post-mortem cortical transcriptomic differential expression analysis between 107 ASD cases and controls reported upregulation of genes associated with microglia activation (M1) and the type I interferon response, with the M2 activation state module negatively correlated with the expression of synaptic transmission genes (222). Rare genetic variants in *CX$_3$CR1* gene, which encodes a fractalkine receptor expressed by myeloid cells, has also been associated with increased risk of ASD (and schizophrenia) (223). In Cx$_3$cr1 deficient mice, excitatory synapses in the hippocampus are structurally immature (188,224,225) and mice exhibit repetitive and reduced social behaviours analogous to ASD phenotypes in humans (188,225), leading authors to speculate that this could be caused by delayed microglial colonisation of the brain during a key neurodevelopmental period (188,226,227).

1.5 Aims

In this thesis, my primary aim was to investigate whether gene regulation in human microglia contributes to genetic risk for complex brain disorders. To do this I measured the open chromatin profile of adult *ex vivo* microglia (chapter 2), *in vitro* human cell models of microglia (chapter 3) and cryopreserved microglia extracted from the 2nd trimester human foetal brain (chapter 4) and integrated these data with GWAS data for complex brain disorders. Finally, using an electrophoresis mobility shift assay, I aimed to investigate how allelic variation at two LOAD risk-associated loci, that overlap adult *ex vivo* microglial open chromatin sites, impacts DNA-protein interactions in microglia (chapter 5).

2 The open chromatin landscape and common variant disease heritability in human *ex vivo* microglia

2.1 Introduction

Although GWASs have been successful in associating multiple genomic loci with risk for complex brain disorders, it is generally unclear in which cells risk variants are active. Given that the majority of brain disorder susceptibility variants fall within intronic or intergenic regions that do not code for protein, risk variants are likely to impact processes that regulate gene expression in disease-relevant cell types (90). As described in chapter 1.3.2, the gene regulatory profile of a cell is partly determined by the biophysical state of chromatin that surrounds specific regulatory elements. Interactions between regulatory elements and transcription factors, that influence gene expression, are permissible in regions where chromatin is open and the DNA is exposed. As such, each cell-type has a unique set of accessible regulatory elements that have the potential to be activated and these elements dictate a cell-type's function and identity. This is important when considering the impact that a risk variant may have in a particular cell type as, if the variant falls within in an inactive region, it can have no impact on the function in that particular cell type. Understanding how individual brain cell types contribute to the burden of risk for complex brain disorders is important to understand their underlying aetiologies.

Chromatin-based assays can identify cell-specific regulatory elements and make it possible, when integrated with GWAS data, to test for enrichment of GWAS risk loci within such annotations. For example, in a recent study by Tansey et al., GWAS summary statistics were integrated with H3K4Me3 ChIP-seq data derived from neurons and glia extracted from human post-mortem tissue (228). The authors demonstrated that, whilst schizophrenia SNP heritability was enriched at promoter sites common to both cell types, when testing for enrichment at promoters that were specific to either cell type, only promoters in neurons were enriched. This implies that neurons contribute more to the common genetic load of schizophrenia than glia in adult brain.

Genetic evidence strongly implicates microglia function in the pathogenesis of late-onset Alzheimer's disease (LOAD). For example, the majority of genes implicated by

LOAD GWAS, including *SPI1*, *CR1*, *CD33*, *ABI3*, and *PLGC2* (79,80,229), are preferentially expressed by microglia, and pathway analyses implicate immune and microglia specific gene networks with increased risk (209). Rare variation in *TREM2*, which encodes an extracellular myeloid cell receptor, is also associated with increased risk for LOAD (229). Moreover, environmental insults that putatively increase microglial activation, such as maternal immune activation, have also been associated with increased risk of schizophrenia and autism (230–234). Despite these findings, the role that microglia have in the aetiologies of these disorders is unknown. As such, I sought to investigate the causal relationship between common variant genetic risk for complex brain disorders and dysfunction in microglial regulatory regions.

2.1.1 Aims

In this chapter, using a statistical technique called stratified linkage disequilibrium score regression (sLDSC), I test whether SNP heritability, measured in a panel of complex brain disorder GWAS, is enriched in genomic open chromatin sites derived from human adult *ex vivo* microglia. I then run a second analysis substituting the microglia data for open chromatin regions derived from human adult neurons to ascertain whether any SNP heritability enrichment I measure in microglia annotations, is specific to microglia or common across brain cell types. Finally, I further partition the microglial annotation into microglial open chromatin sites that contain individual transcription factors to determine whether brain disorder SNP heritability is enriched in open chromatin sites that contain specific factors involved in the regulation of gene expression. This work was included as part of a recent publication (235).

2.2 Methods

2.2.1 Accessing publicly available datasets

ATAC-seq data derived from adult human *ex vivo* microglia, which were extracted from fresh surgically resected brain tissue (236), were obtained from the database of Genotypes and Phenotypes repository (dbGaP) using the study accession code **phs001373.v1.p1**. For comparison, ATAC-Seq data derived from adult human

neurons extracted from frozen ventrolateral prefrontal cortex (135) were downloaded from the European Bioinformatics database quoting the study accession code **PRJINA380200**.

2.2.2 Sequencing, QC and bed file preparation

In total, 12 single-end human *ex vivo* microglia fastq files (corresponding to 12 biological replicates) and 10 paired-end human vlPFC neuronal fastq files (corresponding to 5 biological replicates) were downloaded. The sequencing depth of each file set ranged between 28.3-45.6 million and 20.3-43.6 million reads respectively. For quality control, FastQC (237) was run separately on all fastq files and MultiQC (238) used to collate the FastQC output. As all files passed the quality control measures, no files were excluded from the analyses. Next, single-end and paired-end fastq files were aligned to the human genome using Bowtie2 (239), and the average alignment rates were 98.3% for the microglial files and 99.0% for the neuronal files. Mitochondrial reads were removed from all files using Samtools (240). For peak calling, MACS2 (241) was run using either the `BAM` parameter for single-end files or `BAMPE` parameter to handle paired-end reads, and an FDR of < 0.05 was set as the threshold. Duplicate reads were ignored by MACS2 during the analysis by default. The final output from the peak calling process is a peak file. Diffbind (242) was used to obtain consensus peak files that retained high confidence open chromatin regions for each cell type, with the consensus setting at 0.66, in order to retain peaks that were observed in at least 2/3rds of all donor files for each cell type.

2.2.3 De novo motif enrichment analysis

HOMER motif analysis (243) was used to test enrichment of transcription factor binding motifs in the adult microglia open chromatin peak set. First the microglia peak file was annotated using the `annotatePeaks.pl` command. This provides additional information for each peak, such as whether it occurs over a promoter or an intergenic region, and is required for motif analysis. For the motif analysis, the annotation file was used to run the `findMotifsGenome.pl` command. As HOMER motif analysis is a differential motif discovery algorithm, it first creates a set of background sequences with which to compare the regions in the annotation file. This

background set was selected at random from the specified genome (hg19), and the size (`-size given`) and GC content of these sequences were chosen to match the regions provided in the microglial peak file. Then both the background and microglial open chromatin regions were queried for nucleotide stretches of specified lengths (`-len 14,12,10,8`) and these were tested for enrichment of particular sequences using cumulative binomial distributions. Sequences that are enriched in microglial open chromatin regions were then compared for similarity to a database of known transcription factor motifs and reported if the p-value was < 0.05. Two files are produced by homer, containing the enrichment scores for either known, or *de novo* transcription factor motif locations. As I am interested in the discovery of novel motifs, only the *de novo* results are reported here.

2.2.4 Stratified linkage disequilibrium score regression

Stratified linkage disequilibrium score regression (sLDSC) is a statistical method designed to estimate the proportion of SNP heritability associated with a trait in the population. It also makes it possible to partition the genome by functional category and to test whether SNP heritability associated with a trait is attributable to certain functional categories more than others. SNP heritability in this context is defined as the proportion of genetic variance in a trait explained by all SNPs, as opposed to SNPs that reach the genome-wide threshold for significance. To partition heritability, sLDSC requires a GWAS summary statistics file and an annotation file that contains genomic coordinates for the functional partition. sLDSC exploits the predicted relationship that exists between the association statistics of a set of GWAS index SNPs and local linkage disequilibrium surrounding each index SNP for a polygenic trait, i.e. this relationship predicts that, on average, index SNPs with high LD r^2 scores are more likely to tag SNPs with higher association statistics than SNPs with low r^2 scores. As such, using a linear regression model, sLDSC generates an estimate of the proportion of SNP heritability that is captured by a given set of index SNPs, including any heritability explained by non-assayed SNPs within the haplotype block that each index SNP tags, as well as correcting for systematic biases.

To calculate the GWAS SNP heritability associated within microglial or neuronal open chromatin sites, an annotation file was generated by using bedtools to intersect either the microglial, or neuronal consensus peak file with the SNPs contained in the 1000

genomes reference panel. This assigns a 1 or 0 to each SNP depending on whether or not it overlaps an open chromatin region. Next, using the `lsdc.py` script and `plink` files provided, the latter containing genotype data from 329 European ancestry reference genomes, LD scores for all SNPs with a minor allele frequency >5% were taken by measuring the correlation between SNPs in 1cM windows in the `plink` files. However, only LD scores for SNPs that were located in both the GWAS summary stats file and the annotation file were retained for the next stage in the analysis. Next, again using the `lsdc.py` script, heritability was partitioned by functional category by regressing the χ^2 association statistics for each GWAS index SNP present in the annotation file against each SNP's local LD score. This produced a single regression co-efficient representing the per-SNP heritability for that functional category (note that this assumes uniform per-SNP heritability for all SNPs in the category). Finally, the total SNP heritability for the category (h^2) was calculated by multiplying the regression coefficient by the total number of SNPs in the category and an enrichment score was produced. The significance of SNP heritability enrichment was tested by generating a further 200 SNP heritability regression coefficients from random, equally sized, blocks of SNPs. This provided a normal distribution of regression coefficients and made it possible to calculate a z-score for the SNP heritability coefficient of the functional annotation of interest relative to the normally distributed coefficients. This score includes a correction for the baseline model which includes 53 annotations representing intrinsic heritability that is attributable to common genomic signatures across cell types such as promoters proximal to housekeeping genes, shared chromatin features and evolutionarily conserved regions (244). As these annotations are common between cell types, they would inflate the heritability score if not accounted for in the analysis. Z-scores were transformed and reported as p-values, with significance taken as $p < 0.05$.

2.3 Results

2.3.1 De novo motif enrichment analysis

Table 2.1 shows the results for the *de novo* motif enrichment analysis in human *ex vivo* microglial open chromatin regions. The highest ranked motif was Spi1, which was significantly enriched in *ex vivo* microglial open chromatin sites compared to a random set of background regions (target = 32.34%, background = 7.97%; p-value =

1 x 10^{-8874}). Spi1 is a mouse gene that encodes the transcription factor Pu.1. Pu.1 is a myeloid lineage determining transcription factor and both it, and its human homolog PU.1 (that is encoded by the gene *SPI1*) bind to the same transcription factor motif. Pu.1 is critical to microglial development and differentiation such that microglia are absent in mice lacking the Spi1 gene (146).

Table 2.1. Motif enrichment analysis in human *ex vivo* microglia OCRs

Rank	Motif	% of Targets	% of Background	p-value
1	Spi1	32.34	7.97	1 x 10-8874
2	BORIS	11.62	1.58	1 x 10-4950
3	IRF8	9.45	2.37	1 x 10-2271
4	Sp5	13.58	8.08	1 x 10-623
5	CEBPE	12.43	7.25	1 x 10-606
6	RUNX	24.77	17.66	1 x 10-578
7	BATF	8.23	4.43	1 x 10-499
8	DCE	21.19	15.06	1 x 10-484
9	E2F2	21.18	15.99	1 x 10-337
10	nMyc	13.59	9.41	1 x 10-333
11	NFY	10.93	7.28	1 x 10-315
12	MafA	24.66	19.41	1 x 10-301
13	ETV5	22.33	17.21	1 x 10-300
14	CREB1	7.88	5.2	1 x 10-233
15	NRF1	5.77	3.6	1 x 10-209
16	IRF5	4.2	2.43	1 x 10-198
17	USF2	5.06	3.14	1 x 10-188
18	MEF2A	3.21	1.87	1 x 10-147
19	GFY	1.31	0.58	1 x 10 -123
20	Rfx1	1.31	0.61	1 x 10 -112

% of Targets (proportion of microglial open chromatin sites containing specified motif); % of Background (proportion of randomly selected backgroun regions containing specifed motif); OCRs (open chromatin regions).

Moreover, both interferon regulatory factor 8 (IRF8) and Runt-related transcription factor 1 (RUNX), the third and sixth ranked motifs respectively, were significantly enriched in microglial open chromatin regions over the background (IRF8, target = 9.45%, background = 2.37%, p-value = 1 x 10^{-2271}; RUNX, target = 24.77%, background = 17.66%, p-value = 1 x 10^{-578}). Both factors are critical regulators of microglial differentiation (144,146,245). Therefore, the transcription factor motif profile of the adult *ex vivo* microglial open chromatin regions is consistent with what one would expect for a functional annotation derived from a myeloid cell.

2.3.2 Enrichment of brain disorder GWAS SNP heritability in human *ex vivo* microglial open chromatin regions

To test for enrichment of SNP heritability in *ex vivo* microglial open chromatin sites, sLDSC was run using GWAS SNPs from 7 brain disorder traits: attention deficit hyperactivity disorder (246) Alzheimer's disease (80), autism spectrum disorder (83), bipolar disorder (84), major depressive disorder (85), neuroticism (247), and schizophrenia (248). A GWAS of wearer of glasses or contact lenses (WGL; downloaded from http://www.nealelab.is/uk-biobank/) was used as a negative control.

Table 2.2. Enrichment of brain disorder GWAS SNP heritability in human *ex vivo* microglial OCRs

GWAS	SNPs (%)	h2 (%)	h2 SE (%)	Enrichment	Enrichment SE	p-value	cor. p-value
ADHD	1.17	4.92	4.25	4.21	3.64	0.492	1.000
ASD	1.17	2.82	5.27	2.42	4.51	0.620	1.000
BPD	1.17	4.03	3.32	3.45	2.84	0.936	1.000
LOAD	1.17	50.47	17.24	43.20	14.76	**0.013**	0.104
MDD	1.17	4.11	4.10	3.53	3.51	0.998	1.000
NEUROTICISM	1.17	-1.01	1.80	-0.87	1.53	0.097	0.776
SCZ	1.17	2.59	2.10	2.22	1.75	0.977	1.000
WGL	1.17	12.49	8.89	10.69	7.61	0.347	1.000

SNPs (proportion of GWAS SNPs contained in SNP reference panel); h2 (SNP heritability); h2 SE (SNP heritability standard error); Enrichment (enrichment of SNP heritability in microglial OCRs); Enrichment SE (enrichment of SNP heritability standard error); cor. p-value (Bonferonni corrected p-value for 8 tests); OCRs (open chromatin regions); ADHD (attention deficit hypractivity disorder); ASD (autism spectrum disorder); BPD (biploar disorder); LOAD (late onset Alzheimer's Disease); MDD (major depressive disorder); SCZ (schizophrenia); WGL (Wearing glasses or lenses)

SNP heritability associated with LOAD was significantly enriched in *ex vivo* microglial open chromatin regions (P = 0.013), but this did not survive Bonferroni correction for the 8 tests run (see table 2.2). SNPs within these sites captured 50.47% of the total common variant heritability explained by the Alzheimer's disease GWAS SNPs which is an enrichment >43 times over and above the heritability explained by the sLDSC background annotations. By contrast GWAS SNP heritability associated with the remaining 6 complex brain disorders, and the negative control, was not significantly enriched in microglial open chromatin sites.

2.3.3 Enrichment of brain disorder GWAS SNP heritability in human ventrolateral prefrontal cortex neuronal open chromatin regions

Table 2.3 shows the sLSDC results testing for enrichment of SNP heritability associated for the 7 brain disorder GWASs, and 1 negative control GWAS, in open chromatin regions derived from adult ventrolateral prefrontal cortical neurons.

Table 2.3. Enrichment of brain disorder GWAS SNP heritability in human vlPFC neuronal OCRs

GWAS	SNPs (%)	h2 (%)	h2 SE (%)	Enrichment	Enrichment SE	p-value	cor. p-value
ADHD	1.01	8.34	4.04	8.24	3.98	0.103	0.824
ASD	1.01	10.10	5.22	10.04	5.16	0.100	0.800
BPD	1.01	9.53	3.01	9.42	2.98	**0.024**	0.192
LOAD	1.01	-10.20	12.70	-10.07	12.54	0.096	0.768
MDD	1.01	2.70	3.45	2.67	3.41	0.660	1.000
NEUROTICISM	1.01	7.98	1.85	7.88	1.83	**2.90 x 10-3**	**0.023**
SCZ	1.01	9.67	1.74	9.54	1.72	**5.59 x 10-5**	**4.47 x 10-4**
WGL	1.01	8.57	7.00	8.47	6.91	0.544	1.000

SNPs (proportion of GWAS SNPs contained in SNP reference panel); h2 (SNP heritability); h2 SE (SNP heritability standard error); Enrichment (enrichment of SNP heritability in microglial OCRs); Enrichment SE (enrichment of SNP heritability standard error); cor. p-value (Bonferonni corrected p-value for 8 tests); OCRs (open chromatin regions); ADHD (attention deficit hypractivity disorder); ASD (autism spectrum disorder); BPD (biploar disorder); LOAD (late onset Alzheimer's Disease); MDD (major depressive disorder); SCZ (schizophrenia); WGL (Wearing glasses or lenses)

In contrast to the *ex vivo* microglial sLDSC results, SNP heritability associated with LOAD was not significantly enriched at the $P < 0.05$ threshold in neuronal open chromatin regions ($P = 0.096$, enrichment = -10.07). However, SNP heritability for bipolar disorder ($P = 0.024$, enrichment = 9.42), schizophrenia ($P = 5.59 \times 10^{-5}$, enrichment = 9.67) and neuroticism ($P = 2.90 \times 10^{-3}$, enrichment = 7.88) were significantly enriched within these regions, with enrichment of SNP heritability for schizophrenia and neuroticism remaining significant after Bonferroni correction for 8 tests.

2.3.4 Enrichment of Alzheimer's disease GWAS SNP heritability in human *ex vivo* microglial open chromatin regions containing specific transcription factors

Due to such large proportion of Alzheimer's disease SNP heritability being attributable to open chromatin regions in *ex vivo* microglia (<50%), see section 2.3.2, I tested whether this heritability could be partitioned further, i.e. to open chromatin regions

containing specific transcription factor motifs. For each of the 20 motifs identified in the *de novo* motif enrichment analysis (see section 2.3.1), a separate peak file was generated that included only those peaks in the original *ex vivo* microglia peak file that contained the specific motif. A peak file comprising of open chromatin sites that contained none of the 20 motifs was also generated as a negative control. The results are shown in table 2.4.

Table 2.4. Enrichment of LOAD GWAS SNP heritability in human *ex vivo* microglial OCRs containing specific transcription factors

Motif	h2 (%)	h2 SE (%)	Enrichment	Enrichment SE	p-value
Spi1	29.08	11.94	65.49	28.90	0.013
BORIS	7.60	6.95	63.29	57.88	0.401
IRF8	4.67	4.58	34.33	33.74	0.411
Sp5	3.07	7.08	12.96	29.90	0.834
CEBP-ε	13.80	6.12	68.74	30.47	0.035
RUNX	20.67	8.30	55.43	22.27	0.033
BATF	5.70	5.10	47.29	42.30	0.352
DCE	11.76	7.70	33.87	22.14	0.423
E2F2	7.18	8.71	20.51	24.91	0.89
nMyc	3.73	5.82	15.61	24.41	0.997
NFY	-2.19	4.65	-12.46	26.44	0.317
MafA	10.64	7.69	29.14	21.05	0.418
ETV5	9.25	7.63	24.97	20.60	0.704
CREB1	13.50	6.00	104.91	46.60	0.256
NRF1	1.30	5.09	11.49	44.76	0.848
IRF5	5.95	3.59	87.84	52.95	0.114
USF2	5.72	3.91	65.94	45.00	0.22
MEF2A	5.98	3.57	134.13	80.09	0.091
GFY	1.82	1.32	8.24	60.29	0.856
Rfx1	2.10	1.65	113.93	89.77	0.263
No TF	1.44	2.91	32.91	66.59	0.708

h2 (SNP heritability); h2 SE (SNP heritability standard error); Enrichment (Enrichment of SNP heritability in micoglial OCRs); Enrichment SE (Enrichment of SNP heritability standard error); OCRs (open chromatin regions); LOAD (late onset Alzheimer's disease)

LOAD SNP heritability was significantly enriched in *ex vivo* microglial open chromatin regions containing one of 3 transcription factor motifs at the $P < 0.05$ threshold; namely, Spi1 ($P = 0.013$), CEBP-ε ($P = 0.035$) and RUNX ($P = 0.033$). Open chromatin sites containing these factors explained 29.08%, 13.80% and 20.67% of LOAD common variant heritability, which corresponds to 65-fold, 69-fold and 55-fold enrichments, respectively. No other set of transcription factor containing open chromatin sites was significantly enriched for Alzheimer's disease SNP heritability at

the P < 0.05 threshold, including the negative control set that contained no transcription factor motifs (No TF). No set of microglial open chromatin sites tested were significantly enriched for LOAD SNP heritability after Bonferroni correction for 21 tests.

2.4 Discussion

Establishing how common genetic risk variation impacts individual cell types in the brain is vital to improve our understanding of the causal mechanisms of complex brain disorders (90).

The localisation of >50% of the total LOAD SNP heritability to adult *ex vivo* microglial open chromatin regions shown here (see table 2.2) implies that microglial specific gene regulatory processes mediate LOAD genetic risk mechanisms in some way, and that these processes account for more than half of the common variant liability for the disease. Moreover, given that the LOAD risk signal was not replicated in neuronal open chromatin sites (see table 2.3), these results add to growing body of evidence that the innate immune system and, myeloid cells specifically, are involved in LOAD pathophysiology. For example, LOAD SNP heritability has been shown to be enriched in myeloid cells extracted from peripheral blood (monocytes and macrophages; 246), and LOAD susceptibility alleles have been reported to be enriched for monocyte-specific cis-eQTL effects (251). The results presented here are the first to directly link LOAD common variation to gene regulatory function in microglia (the primary myeloid cell of the brain), however, as microglia and peripheral blood myeloid cells share many gene regulatory features, it is still unclear whether microglia are the key disease mediating cell-type or whether peripheral myeloid cells contribute in some way.

Having established that SNP heritability was enriched in microglial open chromatin sites, I investigated whether this enrichment could be localised to open chromatin regions containing motifs for specific transcription factors. Whilst distinct cell types share many transcriptional regulators that control common cellular processes, a significant proportion of the gene regulatory landscape is distinguishable between cell types, particularly the enhancer landscape (157,252). In the motif-specific sLDSC analysis, LOAD SNP heritability was significantly enriched in open chromatin sites containing motifs for 3 transcription factors (see table 2.4). All 3 of these transcription

factors are prominently expressed in myeloid cells (245,253,254). For example, Spi1 is a haematopoietic pioneer factor and it has been shown to determine both the constitutive and signal specific enhancer landscape in macrophages (microglia are the macrophage of the brain; 251,252). Common missense mutations in the *SPI1* gene are associated with LOAD (79), and PU.1, the transcription factor encoded by *SPI1*, has been reported to be present in the cis-regulatory elements of LOAD-associated genes in myeloid cells (monocytes and macrophages), suggesting that SPI1 modulates LOAD risk via regulation of LOAD-associated genes (208).

Interestingly, in a gene expression study carried out by Olmos-Alonso et al. on post-mortem tissue comparing LOAD cases and non-symptomatic controls (257), expression of Spi1, Runx1 and CCAAT-enhancer-binding protein-α (CEBP-α), which binds to same core transcription factor motif as CEBP-ε, were shown to be upregulated in the temporal cortex in LOAD cases. These factors are key regulators of the colony stimulating factor 1 receptor (CSF1R) pro-mitogenic pathway that acts to increase microglial proliferation which, the authors show, positively correlates with progression of the disease. Olmos-Alonso et al., also demonstrate that inhibition of the CSF1R in an Alzheimer's disease mouse model (APPsew/PSEN1dE9) leads to a reduction in microglial proliferation, a reduction in expression of all three factors (Spi1, Runx1 and Cebp-α) and, crucially, short-term memory recovery and improved performance in exploratory and problem solving tasks designed to measure AD-like cognitive deficits (257). As such, the results presented in this chapter links polygenic common variant risk associated with LOAD to microglial specific gene regulatory networks that have potential therapeutic validity. Moreover, from a methodological perspective, the data presented here demonstrates the utility in integrating GWAS summary statistics with cell specific functional annotations as a means to draw meaningful, and more precise, conclusions about common variation in a cellular context that is not possible using GWAS data alone.

Given that microglia function has been suggested to play a role in several psychiatric disorders (258,259), the lack of SNP heritability enrichment in microglial open chromatin sites for all psychiatric disorders included in the analysis (see table 2.2) is of interest. If microglia do have a role in the neuropathology of these disorders, these results imply that this will be generally mediated by factors other than common genetic variation. Infection and chronic stress during critical neurodevelopmental periods, both of which stimulate immune, and microglial activation, increase risk for

schizophrenia and autism (258) suggesting that environmental effects on microglia may be important in the aetiology of these disorders. In addition, rare genetic variants in the *CX_3CR1* gene, which encodes a fractalkine receptor expressed by myeloid cells, has also been associated with increased risk of ASD (and schizophrenia) (223).

There were several limitations to this study. Firstly, the tissue preparation protocols for the microglial and neuronal datasets differed. For example, *ex vivo* microglia were extracted from live patients and processed within a few hours of extraction (157,236) whereas neurons were extracted from frozen post-mortem tissue obtained from brain banks (135). As these data derived from publicly available sources, it was not possible to control for any confounding effects caused by differing protocols that may have impacted the results presented (260). Secondly, the SNP heritability estimates calculated by sLDSC are based on an additive model. This assumes that the effects of all observed SNPs contribute to heritability in an additive manner, and that each SNP effect is independent from the effects of all other SNPs. Whilst this may be the case for a majority of risk SNPs, it is an oversimplification as it does not take into account of non-additive, epistatic interactions between SNPs which are difficult to measure accurately (244). Thirdly, although the proportion of SNP heritability attributed to each microglial open chromatin annotation is relatively large (>50% and ~14-29% in non-transcription factor, and transcription factor specific open chromatin sites respectively), it should be noted that the SNP heritability captured by sLDSC explains only 6.1% of total heritable component of LOAD. Whilst this may indicate that common variation makes only a small contribution to the overall genetic liability for LOAD, it is likely, at least partly, an artefact of the relatively low sample size of the LOAD GWAS (80) compared to other brain disorder GWAS such as schizophrenia (82). Fourthly, it has been reported that there is a correlation between lowering the proportion of SNPs included in an LDSC analysis and a reduction in the accuracy of its results (244,261). As such, it is not recommended to run sLDSC when <1% of the reference panel of SNPs is represented in the functional annotation of interest (as this may result in sLDSC model misspecification; https://groups.google.com/forum/#!topic/ldsc_users/Mm0zN8ijsiE). As the total proportion of reference panel SNPs included in the microglial and neuronal analyses were 1.17% and 1.01% respectively (see tables 2.2 and 2.3), which is approaching the recommended threshold for the sLDSC model, these results should be interpreted with caution. However, as the proportion of SNPs included in an sLDSC analysis is a function of the genomic coverage of the cell-specific functional annotation, i.e. the

open chromatin regions of the microglial/neuronal genome, the number of SNPs included in the sLDSC analyses was outwith my control.

2.4.1 Future Work

Whilst this work provides better cellular and molecular resolution with respect to GWAS risk loci, it does not provide direct evidence showing whether or not transcription factors are bound at these locations, and more importantly whether the binding of these factors is actually altered by common risk alleles. Future work should seek to address these questions. For example, computational techniques exist to measure occupancy at transcription factor motifs using ATAC-seq data (262). These methods detect dips in read coverage in open chromatin regions of relatively high read coverage, called transcription factor footprints, which represent the locations within peaks where transcription factors are bound. This occurs as the tn5 transposase enzyme normally cleaves DNA (relatively) indiscriminately throughout the entire nucleosome free region but is prevented from doing so when a transcription factor is present. Consequently, there is a reduction of read ends being cut at that locus. As transcription factor footprinting is sensitive to the depth at which ATAC-seq libraries are sequenced, this type of analysis was not possible in this study using the Gosselin et al. *ex vivo* microglial data (236) as the data were not sequenced deeply enough. As such, future analyses should consider increasing the sequencing depth to make transcription factor footprinting possible. ChIP-seq data could also be generated to measure genome-wide occupancy of specific transcription factors in microglia. However, this technique requires a relatively high number of input cells; therefore, due to the poor availability of primary microglia, ChiP-seq data derived from microglia are currently only available for one human transcription factor (*SPI1;* 235).

Given that microglia are a phenotypically dynamic cell type, and that phenotype-specific gene expression may require the open chromatin landscape of microglia to change, it is possible that GWAS risk loci operate in specific cellular contexts, such as particular microglial activation states. In a recent study measuring the open chromatin landscapes of T-cells after exposure to 13 different cytokine cocktails *in-vitro* (which elicit specific immune related activation states), autoimmune GWAS SNPs were shown to be enriched in open chromatin sites in T-cells polarised toward a specific cellular state (263). For example, inflammatory bowel syndrome associated GWAS SNPs were in enriched in open chromatin regions of T-cells skewed toward a

Th1 phenotype, whereas arthritis associated risk SNPs were enriched in open chromatin regions of T-cells pushed toward the Th2 phenotype. A similar analysis assessing the open chromatin landscape of microglia across differing activation states and testing for enrichment of LOAD GWAS SNPs across open chromatin sites in these distinctly activated cells, could provide insight into the specific microglial functions that underpin LOAD aetiology.

2.4.2 Concluding remarks

By integrating brain disorder genetic association data with functional annotation data from a specific brain cell type, I provide evidence that a substantial proportion of common variant genetic risk for LOAD operates through gene regulatory processes in microglia, thereby strengthening the case that microglia are involved in LOAD pathophysiology. Furthermore, I provide evidence that variants contributing to polygenic risk of LOAD are enriched within open chromatin sites containing specific microglial transcription factor binding sites, thus suggesting tangible molecular targets for future mechanistic and/or translational studies.

3 The open chromatin landscape of *in-vitro* human cell models of microglia

3.1 Introduction

Access to fresh human brain tissue for brain disorder research is extremely limited. As such, procurement of human tissue mainly derives from frozen post-mortem sources (264). Brain samples that can be obtained from living patients tend to derive from individuals with existing brain pathology, such as epilepsy or tumour, so there is a risk that data generated using such tissue will be confounded by the effects of disease. When fresh tissue is available, it is possible to extract primary microglia for analysis; however, they downregulate many microglial-specific genes within a few hours of being removed from the brain (236,265,266), which severely limits their utility. As such, much of our knowledge regarding human microglial function is based on data that have been extrapolated from animal models.

Animal models have provided exceptional insight into specific aspects of microglial function in health and disease. For example, the use of transgenic mice made it possible to map the fate of early microglial precursors during early development, establishing the embryonic yolk-sac as the location of microglia origin (144,146), and determining the critical factors that drive microglia differentiation and maturation in mice (i.e. PU.1, Irf8 and IL-34; 145,146,267). This work has directly informed the differentiation strategies used to derive microglia-like cells from induced human pluripotent stem cells (268). Similarly, mouse-derived immortalised microglial lines, such as BV2 cells, have been used to study neurodegenerative disease. For example, Aβ uptake has been shown to be impaired when BV2 cells are transfected to overexpress the LOAD risk gene *CD33* (which has increased expression in LOAD patients), implying that CD33 mediates microglial clearance of Aβ, and that this mechanism may be relevant to AD risk in humans (269).

Immortalised microglial cell lines have several benefits over primary microglia, including that they are freely accessible, easily maintained and propagate in an unrestricted manner (270). However, the legitimacy in using animal-derived immortalised microglial lines has been questioned for several reasons. First, as the immortalisation process often employs a retrovirus to introduce oncogenes to the

microglia genome (271), differences in cell morphology and adhesion have been reported when comparing immortalised lines to primary microglia (272). Indeed, the highly proliferative nature of immortalised lines does not accurately represent the microglial *in vivo* state (268). Secondly, immortalised microglia have a tendency to dedifferentiate and lose microglial specific features (270,273). For example, BV2 cells have been shown to have functional differences when compared to primary and *ex vivo* human microglia (265,274,275). For these reasons, there is a critical need to establish a reliable and renewable human microglial cell model.

One area of promise in relation to the generation of human-derived cell lines for use in the study of brain disorders is in the implementation of induced pluripotent stem cell (iPSCs) technology. iPSCs are generated from somatic cells, such as skin cells (fibroblasts), that are reprogrammed to express genes which elicit a pluripotent state in the cells. Then, using a cocktail of factors and cytokines pluripotent cells can then be skewed toward a specific cellular fate (276). Compared to primary brain cells, iPSC lines provide similar benefits to animal cell lines, such as their accessibility and capacity to be cultured for extended periods; however, a key additional benefit in relation to complex brain disorders is that iPSC lines can be derived from patients with disease-specific genetic backgrounds. This makes it possible to compare cellular phenotypes between affected cases and control individuals and use gene editing technology in these cells to test hypotheses regarding the genes and molecular mechanisms predicted to drive these disorders. Several protocols have been published recently describing how to generate microglia from induced pluripotent stem cells (iPSCs) derived from human fibroblasts (268,277–279). Whilst the precise methodologies used across these studies differ in terms of the specific cytokines and incubation times used to generate iPSC-derived microglia (iPSC_MG), and whether or not additional brain cell-types were added in co-culture, they all claim to generate viable myeloid cells with microglia-like properties. However, as transcriptomics was used as the primary global indicator to measure iPSC_MG in all of these studies, it has yet to be established whether the open chromatin landscape of microglia derived from iPSCs recapitulates that of primary microglia, despite the fact that a cell's enhancer landscape better predicts cell identity than its transcriptomic profile (252).

3.1.1 Aims

In the work described in this chapter, I aimed to generate three novel ATAC-seq datasets from human cell culture lines. The first two lines were iPSCs harvested at distinct developmental time points during the myeloid iPSC differentiation process. These lines were iPSC macrophage precursors (iPSC_MΩpre), which are differentially antecedent to microglia and the terminally differentiated iPSC microglia (iPSC_MG). The third cell line was a human microglial line immortalised using the simian virus 40 (SV40). Motif enrichment and principal component analyses were then used to compare the open chromatin landscapes of all three cell lines to the landscapes of human *ex vivo* cells of myeloid and lymphoid lineage taken from the brain and circulatory system. Primarily, I aimed to establish whether the iPSC_MG microglial cell line can adopt a chromatin accessibility profile similar to that seen in human *ex vivo* microglia.

3.2 Methods

3.2.1 Accessing publicly available datasets

To compare the open chromatin landscapes of the iPSC lines to microglia *ex vivo* open chromatin sites, the chromatin accessibility data from the Gosselin and colleagues study was again used (see section 2.2.1 for repository and cell preparation details; 280). A second *in vitro* microglial open chromatin dataset from the same repository was also included. These *in vitro* cells were derived from the *ex vivo* microglial population as described in section 2.2.1 but, before ATAC-seq took place (280), the cells were first cultured and maintained in medium (Dulbecco's Modified Eagle Medium with 5% FBS and 20ng/ml Interleukin-34) for 7 days to assess the impact of environmental conditions on microglia phenotype. Additionally, due to the functional similarity of myeloid cells deriving from the brain and peripheral circulatory system, chromatin accessibly datasets extracted from a suite of peripheral blood cell populations (myeloid and lymphoid), were obtained for comparison (252). Data were produced for the peripheral blood datasets by Corces and colleagues by blood extraction from patients and separated into its individual components using Ficoll to create a cell density gradient under centrifugation. Cells were then cryopreserved (90% FBS + 10% DMSO) in liquid nitrogen until required for ATAC-seq. The cell types

included in the analysis were monocytes, CD8 T cells, CD4 T cells, natural killer cells and B cells. These data were downloaded from the GEO via accession code **GSE74912**.

3.2.2 Processing induced pluripotent stem cells

Undifferentiated iPSCs were generated from the commercially available Kolf2 cell line (HipSci - http://www.hipsci.org/). These iPSCs derived from dermal fibroblasts taken from a male of European descent aged between 55-59 years. The iPSC differentiation procedure was undertaken by my colleague Aurelian Bunga using the protocol published by Haenseler and colleagues with slight modifications (268). A brief description of the iPSC_MΩpre and iPSC_MG differentiation protocol he used is as follows. To promote the formation of embryoid bodies, ~3 million human-derived iPSCs were seeded into plate wells and covered with mTeSR1 media supplemented daily with bone morphogenic protein 4, vascular endothelial growth factor and stem cell factor. After 4 days, embryoid bodies were re-plated and incubated in X-VIVO15 media with macrophage colony stimulating factor (M-CSF), interleukin-3, glutamax, penicillin, streptomycin and β-mercaptoethanol, which was replenished every 7 days. After ~1-month, iPSC_MΩpre emerged. To obtain iPSC_MG, the iPSC_MΩpre were incubated in advanced DMEM/F12 supplemented with N2, granulocyte/macrophage colony stimulating factor (GM-CSF) and interleukin-34 (IL-34) for 7-14 days. Both iPSC_MΩpre and iPSC_MG were extracted at their appropriate differentiation stages, strained, cryopreserved in Hibernate E (Gibco) and stored at -80^{o}C until required for ATAC-seq.

3.2.3 Immortalised human microglia – SV40

Immortalised microglial cell lines are widely available for a range of species including human. The benefits of using an immortalised line it is easier to culture than an iPSC, or primary culture line (as it has been modified to evade normal cellular senescence) and it provides an unlimited and relatively consistent cell source. The human microglia SV40 cell line (SV40s; Applied Biological Materials) was used in this study to test whether its open chromatin landscape was similar to that measured in *ex vivo* microglia. SV40s were derived from primary human microglia and immortalised by

transduction and serial passaging with recombinant lentiviruses carrying simian virus 40 large T antigen (281).

The cell processing procedure for the SV40s was as follows. When cell seeding was required, 1-2 cryovials of SV40s were obtained from in-house stocks cryopreserved in liquid nitrogen at -180°C. To defrost the cells, the cryovials were placed in a 37°C water bath for ~2 minutes. All cryopreserved SV40s were of low passage number (2-5 passages). For the general cell maintenance and passaging, T75 flasks were pre-coated with collagen (5µg of collagen per cm^2) then incubated for 60 minutes at room temperature. The collagen was then removed, and the flasks washed 3 times with phosphate-buffered saline (PBS). Defrosted cells were then carefully added to the collagen coated flasks and incubated in Pirigrow III (Applied Biological Materials), supplemented with 10% fetal bovine serum (Sigma) and 2mM L-glutamine (Sigma), with a 5% CO_2 level and the temperature set at 37°C. Cell culture media was changed every 2-3 days and cells were passaged when 80-100% confluent. First the media was extracted from the flasks and, to dissociate the cells, 3ml of Accutase (ThermoFisher) was added to each flask, gently swirled to cover all cells, then flasks were incubated for 3 minutes at 37°C. Once the cells had detached, 3ml of media was added to neutralise the Accutase and the cell suspension was centrifuged at 1000rpm in a 15ml Falcon tube for 1 minute. The supernatant was removed, and the cell pellet was resuspended in the desired volume of supplemented Pirigrow III media. Finally, the cell suspension was seeded into new collagen coated flasks. This process was repeated until cells were required for ATAC-seq.

3.2.4 ATAC-seq library preparation

Cells were pelleted (SV40s) as described in section 3.2.3, or defrosted (iPSC_MG and iPSC_MΩpre) in a water bath at 37°C for 2 minutes, and then pelleted, before being transferred into Eppendorf tubes. The ATAC-seq library preparation process was identical for each of the SV40, iPSC_MG and iPSC_MΩpre cell lines and followed the OMNI-ATAC protocol (282). Nuclei were isolated by resuspending the cell pellets in 50µl cold cell lysis buffer (0.1% IGEPAL (Sigma), 0.1% Tween-20 (Sigma), and 0.01% digitonin (Sigma)). Isolated nuclei were washed in wash buffer (0.1% Tween-20), pelleted then resuspended in transposase solution (25ul 2x TD buffer, 2.5ul Tn5 transposase (100nM final; Illumina), 16.5ul PBS, 0.5ul 1% digitonin,

0.5ul 10% Tween-20, 5ul H_2O) and incubated for 30 minutes at 37°C. The DNA library was isolated and cleaned, using the DNA Clean and Concentrator-5 Kit (Zymo) and snap frozen at -20°C in 20µl of elution buffer. When I was ready for the next stage in the process, DNA was defrosted on the bench and amplified for 5 cycles of PCR (Bio-rad S-1000) using the parameters outlined in table 3.1. Importantly, during all PCR and qPCR stages, the primers used for each sample consisted of a generic forward primer (Ad1) and a unique reverse primer (Ad2). These were kept consistent for each sample throughout the protocol.

Table 3.1. PCR reaction protocol

50µl reaction		Cycling conditions	
PCR amplification	**Volume (µl)**	**Temp (C)**	**Time**
25µM Primer Ad1 (Nextera)	2.5	72	5min
25µM Primer Ad2 (Nextera)	2.5	98	30sec
Master mix (2x NEBNext)	25	**Then 5 cycles of:**	
Transposed DNA	20	98	10sec
		63	30sec
		72	1min
		Hold at 4 (C)	

As the total number of amplification cycles required was sample dependant, in order to calculate the number of additional PCR amplification cycles required to generate an adequate DNA library for each sample, 5µl of each pre-amplified sample was aliquoted from the PCR products and added to a 30-cycle qPCR reaction (Applied Biosystems StepOnePlus) using the parameters described in table 3.2.

Table 3.2. qPCR reaction protocol

20µl reaction		Cycling conditions	
qPCR amplification	**Volume (µl)**	**Temp (C)**	**Time**
25µM Primer Ad1 (Nextera)	3.75	98	30sec
25µM Primer Ad2 (Nextera)	0.5	**The 5 cycles of:**	
Master mix (2x NEBNext)	0.5	98	10sec
25x SYBR Green in DMSO (ThermoFisher)	0.25	63	30sec
Pre-amplified sample	5	72	1min
18.2Ω water	5	**Hold at 4 (C)**	

To ascertain the number of additional cycles of PCR required for each sample, a graph depicting the relative fluorescence intensity of each qPCR reaction against cell cycle number was generated and the number of cycles it took for the qPCR reaction to reach 1/3 of the maximum relative fluorescence intensity was calculated. This number corresponds to the number of additional PCR cycles required for each sample, which was between 5-6 cycles. DNA fragments were then isolated and cleaned, as described above, using the DNA Clean and Concentrator-5 Kit (Zymo).

In order to extract DNA fragments of a specific length, representing transposase-mediated insertion of sequencing adapters into either one or two nucleosomes, DNA was size-separated by gel electrophoresis. A 2% agarose solution was made by dissolving agarose (Sigma) in 0.5X TBE (Tris-borate-EDTA; Thermo Fisher), and then ethidium bromide was added (1:100) to the solution. Ethidium bromide makes it possible to visualise DNA when it is exposed to ultraviolet light. The solution was then heated to until boiling point, and carefully added to a sealed gel-cast. A gel-comb was added across the cast to create wells in the gel, and the gel was left to set for 60 minutes. Once the gel had set, a 100bp DNA ladder (New England Biolabs) was added to the right- and left-most wells of the gel. PCR products, pre-mixed with 10X loading dye (Orange G) were then added to the gel, one sample per well, leaving a gap of one well between all ladder and sample wells to prevent cross-contamination. The gel was electrophoresed at 100V for 20-25 minutes such that the increments in the DNA ladder had separated sufficiently to identify DNA between the range of 175-400bp. Amplicons of this size range were selected using a scalpel under ultraviolet light and DNA extracted from the gel using the Zymoclean^{TM} Gel DNA Recovery Kit (Zymo).

To prepare for sequencing, the ATAC-Seq library was quantified using the Qubit 2.0 Fluorometer (range 0.48-1.04ng/µl) and the average fragment length of the DNA in each sample was measured (range 307bp) using the Agilent high sensitivity DNA chip. DNA was pooled to a final molarity of 10nM in 30µl of elution buffer (Zymo) sequenced using an in-house Illumina HiSeq 4000 system (sequencing performed by Dr Joanne Morgan).

3.2.5 Sequencing, QC and bed file preparation

Paired-end 75-base sequencing was carried out on the 3 SV40, and 6 iPSC cell line ATAC-seq libraries using the Illumina HiSeq 4000. Sequencing files were de-multiplexed then sequencing adapters and poor-quality reads were removed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), resulting in 18 cleaned fastq files which contained either the forward or reverse reads for one the 9 ATAC-seq libraries sequenced. In addition, a total of 15 single end and 52 paired end ATAC-seq fastq files were downloaded from the public repositories respectively (see section 3.2.1). These corresponded to a panel of open chromatin datasets derived from cells of the brain and circulatory systems. In total, data from 7 additional cell-types were included in the analysis.

Table 3.3. Cell type specific open chromatin data included in the chapter 2 analyses

Cell Type	No. of Replicates	Replicate Type	Read Depth (M)	Alignment Rate (%)	Peak No. (K)	Study
SV40s	3	Technical	104.1 - 140.8	98.4	169.6	This study
iPSC MG	3	"	109.4 - 173.0	97.1	211.1	"
iPSC MΩpre	3	"	202.0 - 360.2	97.1	249.6	"
MG ex-vivo	12	Biological	28.3 - 45.6	98.3	100.6	Gosselin et al., (2017)
MG in-vitro	3	'	32.2 - 51.4	98.2	122.3	"
CD8	5	Biological	30.6 - 122.8	62.8	33.2	Coerces et al., (2016)
CD4	5	'	55.0 - 127.4	62.2	34.3	"
NK cells	6	'	27.4 - 124.8	56.8	35.1	"
Bcells	6	'	50.0 - 184.2	65.4	30.8	"
Monocytes	4	'	15.2 - 97.8	71.1	22.1	"

The downstream processing of the fastq files was identical to that described in section 2.2.2. As before, FastQC (237) was used for quality control and MultiQC (238) used to collate the FastQC output. All files passed the quality control measures. Fastq files were aligned to the human genome (hg19), using Bowtie2 (239), and mitochondrial reads were removed from all files using Samtools (240). MACS2 (241), was used for peak calling, duplicate reads were ignored, and the FDR threshold for peak significance was set to <0.05. In total, 50 unique peak files were generated and unique peak files for each cell type were combined to create a consensus peak file using Diffbind (242). The consensus threshold was set to 0.66, in order to retain high confidence peaks that were observed in at least 2/3rds of all donor files for each cell type. The quality measures for all files included in the analysis are summarised in table 3.3. For peak visualisation, technical replicate bam files for each cell type were

merged using Samtools (240) and peaks were visualised using the integrative genome viewer (IGV; 281).

3.2.6 Motif enrichment analysis

Homer motif enrichment analysis (243) was run on the iPSC_MG, iPSC_MΩpre and SV40 cell consensus peak files to ascertain whether motifs enriched in the open chromatin regions of these cell culture lines were similar to the motifs measured in the *ex vivo* microglial open chromatin regions generated in section 2.3.1. The Homer parameters were set exactly as specified in section 2.2.3.

3.2.7 Principal component analysis

Principal component analysis (PCA) is an unsupervised statistical method used to transform multi-dimensional data. It provides a means to summarise complex datasets to identify patterns in the data, or relationships between samples, that may exist but would otherwise be difficult to distinguish (284). To approximate the similarity between the open chromatin landscapes of the 3 cell culture lines I generated and the open chromatin datasets of the *ex vivo* and *in vitro* cells that I downloaded, a PCA was run using the Diffbind package (242). To run the PCA, a file containing the bam file and peak file locations for all 50 samples shown in table 3.3 were produced, which included metadata relating to each sample, such as the cell-types from which the individual datasets were derived and unique replicate identification numbers. Peak files that were included in the analysis were individual peak files for each sample, rather than the consensus peak files. The peak file list was then read into Diffbind using the `dba()` with the `minOverlap` parameter set to `0.66` which ensured that only peaks present in 2/3^{rds} of the individual peak files which derive from the same cell type were retained. Next, a raw count matrix was generated using the `dba.count()` command, with the score parameter set to `DBA_SCORE_READS`. This produces a read count for every peak that was retained in the `dba()` call for every sample, regardless of whether the peak is present in the sample or not. If the peak is not present within a particular sample, it is assigned a read count value of 0. As such, every retained peak across all 50 samples is represented, and assigned a read count value, in each individual sample. Finally, PCA was run, and a plot produced (dba.plotPCA()), using $\log_2$ normalised read counts.

3.3 Results

3.3.1 *De novo* motif enrichment analysis

The iPSC_MG cell line open chromatin regions were significantly enriched for Spi1 (target = 22.35%, background = 7.99%; p-value = 1 x 10^{-6897}), CEBP-α (target = 11.17%, background = 6.65%; p-value = 1 x 10^{-971}) and PU.1-IRF (target = 2.68%, background = 1.29%; p-value = 1 x 10^{-406}) transcription factor binding motifs compared to background sites (see table 3.4). All of these factors are involved in myeloid fate commitment (144,146).

Table 3.4. Motif enrichment analysis in human iPSC microglia OCRs

Rank	Motif	% of Targets	% of Background	p-value
1	Fra1	20.97	6.83	1 x 10-7388
2	Spi1	22.35	7.99	1 x 10-6897
3	BORIS	7.63	1.77	1 x 10-3842
4	Zfx	6.82	3.35	1 x 10-1013
5	CEBPA	11.17	6.65	1 x 10-971
6	EGR2	8.87	5.07	1 x 10-869
7	MED	8.25	4.64	1 x 10-849
8	RUNX2	8.22	4.66	1 x 10-828
9	HINFP	7.74	4.32	1 x 10-812
10	TEAD4	12.98	8.55	1 x 10-771
11	Sp5	14.92	11.18	1 x 10-457
12	PU.1-IRF	2.68	1.29	1 x 10-406
13	DCE	22.55	18.58	1 x 10-349
14	MafA	21.78	17.90	1 x 10-342
15	Ascl2	10.30	7.76	1 x 10-290
16	FOXL1	8.24	6.08	1 x 10-262
17	SpiB	4.14	2.68	1 x 10-248
18	JUND	5.19	3.57	1 x 10-236
19	NFkB-p65	2.53	1.47	1 x 10-222
20	MEF2B	0.91	0.40	1 x 10-171

% of Targets (proportion of microglial open chromatin sites containing specified motif); % of Background (proportion of randomly selected background regions containing specified motif); OCRs (open chromatin regions)

Similarly, in iPSC_MΩpre cell line open chromatin sites, Spi1 (target = 32.83%, background = 11.14%; p-value = 1 x 10^{-12757}), CEBP-α (target = 12.65%, background

= 7.19%; p-value = 1 x 10^{-1417}) and PU.1-IRF (target = 5.88%, background = 2.87%; p-value = 1 x 10^{-962}) motifs were significantly enriched compared to background open chromatin regions (see table 3.5). Crucially, in both iPSC lines Spi1 was either the most, or second most, significantly enriched motif reported, concordant with the *ex vivo* microglia motif enrichment analysis where Spi1 was the most significantly enriched motif.

Table 3.5. Motif enrichment analysis in human iPSC MΩ precursor OCRs

Rank	Motif	% of Targets	% of Background	p-value
1	Spi1	32.83	11.14	1 x 10-12757
2	BORIS	6.23	1.60	1 x 10-3030
3	BATF	11.59	5.64	1 x 10-1979
4	CEBPA	12.65	7.19	1 x 10-1417
5	PU.1-IRF	5.88	2.87	1 x 10-962
6	Sp1	5.41	2.82	1 x 10-745
7	Zfp161	5.39	2.94	1 x 10-646
8	RUNX1	11.74	8.05	1 x 10-624
9	ZNF519	6.07	3.58	1 x 10-574
10	DCE_III	10.05	7.18	1 x 10-424
11	E2F3	4.28	2.63	1 x 10-345
12	MafK	13.66	10.66	1 x 10-338
13	Mef2d	2.31	1.38	1 x 10-200
14	DCE_II	27.33	24.32	1 x 10-184
15	MAFK	12.44	10.33	1 x 10-175
16	GCM1	9.55	7.80	1 x 10-154
17	SPIB	6.86	5.64	1 x 10-102
18	MGA	6.80	5.69	1 x 10-84
19	Nfe212	11.67	10.33	1 x 10-73
20	Zfp809	10.86	9.60	1 x 10-68

% of Targets (proportion of microglial open chromatin sites containing specified motif); % of Background (proportion of randomly selected background regions containing specified motif); OCRs (open chromatin regions)

Table 3.6 shows the *de novo* motif enrichment analysis results for open chromatin regions derived from the SV40 microglia cell line. Of the 20 motifs shown, the highest ranked myeloid cell specific motif was CEBP-α, which was the fifth most significantly enriched motif in SV40 microglial open chromatin sites compared to a background set of open chromatin sites (target = 13.50%, background = 9.86%; p-value = 1 x 10^{-392}).

Table 3.6. Motif enrichment analysis in human SV40 microglia OCRs

Rank	Motif	% of Targets	% of Background	p-value
1	Fra1	27.64	7.06	1 x 10-11345
2	BORIS	8.23	1.85	1 x 10-3550
3	TEAD4	14.82	9.56	1 x 10-807
4	ERG	18.00	12.90	1 x 10-609
5	CEBPA	13.50	9.86	1 x 10-392
6	EBF	21.64	17.15	1 x 10-388
7	Sp5	7.42	4.79	1 x 10-380
8	Zfp161	7.92	5.24	1 x 10-368
9	Zic3	6.34	3.99	1 x 10-357
10	RUNX1	6.22	4.03	1 x 10-311
11	Zic3	15.44	12.23	1 x 10-261
12	IRF1	1.63	0.73	1 x 10-244
13	ETV1	12.67	9.89	1 x 10-233
14	SOX10	21.96	18.68	1 x 10-197
15	Zfp691	27.78	24.26	1 x 10-191
16	c-Jun	6.41	4.68	1 x 10-176
17	BARHL2	20.70	17.69	1 x 10-175
18	Foxo1	20.73	17.74	1 x 10-172
19	Gfi1	7.85	6.02	1 x 10-159
20	NFkB-p65	2.37	1.43	1 x 10-153

% of Targets (proportion of microglial open chromatin sites containing specified motif); % of Background (proportion of randomly selected background regions containing specified motif); OCRs (open chromatin regions)

CEBP-α is a pioneer transcription factor that drives myeloid cell lineage commitment through preferential enhancer selection during myeloid differentiation (285). As for *ex vivo* microglial open chromatin sites (discussed in section 2.3.1), the RUNX1 motif was also significantly enriched in the SV40 open chromatin regions over the background (target = 6.22%, background = 4.03%; p-value = 1 x 10^{-311}), and was the 10^{th} ranked motif reported. However, in contrast to the *ex vivo* microglia motif enrichment results, there was no enrichment for motifs binding the myeloid lineage determining transcription factors Spi1 or IRF8. Indeed, when inspecting peaks over the *SPI1* promoter of the *ex vivo* and *in vitro* microglial cell lines (see figure 3.1), there is a distinct peak present in all lines except the SV40s implying that chromatin is closed and that *SPI1* is likely not expressed in the SV40 line.

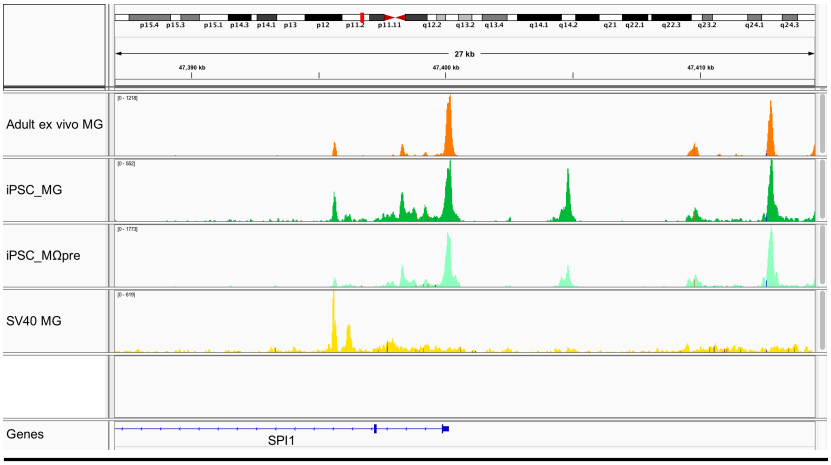

Figure 3.1. ATAC-seq peaks in adult *ex* vivo, and cell lines of, human microglia. Representative genomic tracks showing 27kb surrounding the promoter of the myeloid lineage determining transcription factor *SPI1*. X-axis, genomic location; Y-axis, auto-scaled read counts. Adult ex vivo MG (Adult *ex vivo* microglia); iPSC_MG (induced pluripotent stem cell derived microglia); iPSC_MΩpre (induced pluripotent stem cell derived macrophage precursor); SV40 MG (Simian virus 40 large T antigen immortalised microglia).

Common to all motif analyses is enrichment of generic transcription factor motifs such as Fra1 and BORIS. In humans FRA1 encodes a leucine zipper protein that forms part of the AP-1 transcription factor unit that regulates cell proliferation and differentiation (286) whilst BORIS is a paralog of CTCF, the latter of which encodes an 11 zinc finger protein that can act as a transcriptional activator or repressor (287). As both factors bind the same motif and are expressed in a mutually exclusive manner, (CTCF is widely expressed across tissues whereas BORIS is expresses mainly in gametes) it is likely these motifs bind CTCF (288,289).

3.3.2 Principal component analysis

Figure 3.2 depicts a plot of the first two principal components of the PCA undertaken for the ATAC-seq data from the 50 samples (10 cell types) included in the analysis. The first two principal components explained 85% of the total variance in the data. The first principal component, which captured 77% of the variance, generally separates 8 of the cell lines based on the environmental conditions from which the cells derived, with the trend, when moving from left to right on the plot, shifting from *in-vitro* to *ex vivo* cells. However, two *ex vivo* cell types derived from peripheral blood [i.e. the monocyte (lavender) and B cell (purple) samples] do not follow this trend and cluster with the cells that were processed *in vitro*. This may be

due to the fact these cells were cryopreserved after they were extracted and may have been more sensitive to the cryopreservation process than other *ex vivo* blood cells (NK-cells, CD8 T cells and, CD 4 T cells) processed in the same manner.

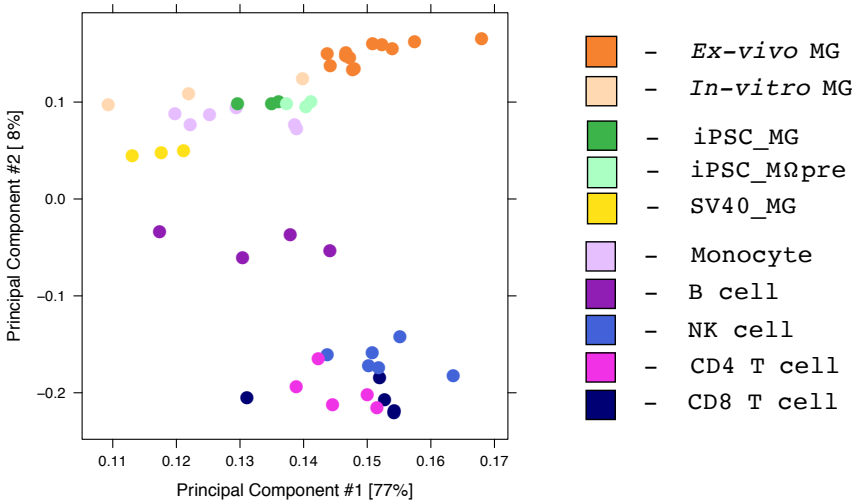

Figure 3.2. Principal component analysis plot of open chromatin site read counts derived from 10 cell lines. *Ex-vivo* MG (*ex-vivo* microglia); *In-vitro* MG (*in-vitro* microglia); iPSC_MG (induced pluripotent stem cell microglia; iPSC_MΩpre (induced pluripotent stem cell macrophage precursors); SV40_MG (SV40 immortalised microglia; Monocyte (blood monocytes); B cell (B cells); NK cell (Natural Killer cells);. X-axis = principal component 1; y-axis = principal component 2. Read counts are $\log_2$ normalised.

The second principal component captures 8% of the variance in the data, and horizontally separates myeloid cells from lymphocytes at 0.0 on the y-axis, indicating that the PCA has clustered cells in a biologically relevant manner and that distinct cell types can be delineated based on differences in their open chromatin profiles.

The iPSC_MG, iPSC_MΩpre and SV40 samples all cluster more closely with the *ex vivo* and *in-vitro* microglial samples and the blood monocyte samples (lavender), than the lymphoid cells samples (B cells, CD4 T cells, CD8 T cells, Mk cells), suggesting that all three cell lines have open chromatin landscapes more similar to myeloid cells than lymphoid cells. Similar to the result in the motif enrichment analysis reported in section 3.3.1, of the three cell lines I generated for the analysis, the open chromatin landscape of the SV40 cells (yellow) appears to recapitulate the landscape of the *ex vivo* microglia cells (orange) the least, as they cluster furthest from the *ex vivo* microglial samples. When comparing the similarity of the iPSC cell lines to the *ex vivo*

microglia, the iPSC_MG cells cluster more closely with the iPSC_MΩpre cells and with monocytes derived from peripheral blood than with the *ex vivo* microglia. This implies that the final stage of the iPSC differentiation protocol, has not been successful in generating a line that recapitulates the open chromatin profile of adult microglia.

3.4 Discussion

Given the difficulty in obtaining primary human brain cells there is an unmet need to manufacture a human-derived cell line that is relatively easy to generate and maintain, and which, as closely as possible, can recapitulate microglial function *in-vivo*. The rationale for this chapter was to measure, and compare, the open chromatin landscapes of 3 human-derived cell culture lines of putative myeloid origin, and to assess how these landscapes corresponded to the open chromatin landscapes of *ex vivo* cells at equivalent stages of myeloid cell development from the brain and circulatory system.

Considering the iPSC lines, the motif enrichment analysis confirmed that these lines do have an open chromatin landscape consistent with what would be expected of a cell from the myeloid lineage. For example, open chromatin sites in both the iPSC_MG and iPSC_MΩpre cell lines contained motifs for key myeloid lineage determining transcription factors such as Sfpi1 and CEBP-α (see tables 3.3 and 3.4), suggesting that the iPSC differentiation protocol is capable of producing cells of myeloid lineage. However, as the motif enrichment profiles of both iPSC lines were similar, and these resembled the profile of *ex vivo* microglia, see table 2.1, it was not possible determine if the iPSC_MG and iPSC_MΩpre were distinct in order to (A) ascertain if the final stage of the iPSC differentiation protocol was successful and that iPSC_MGs did differentiate further to become more microglial-like than iPSC_MΩpre and (B) assess to what extent the open chromatin landscape of the iPSC_MG line resembled that of the *ex vivo* microglial cells. The PCA analysis was performed in an attempt to answer these questions.

As the IPSC_MGs and iPSC_MΩpre samples clustered tightly with one another in the PCA analysis plot (see figure 3.2) and also cluster more closely with blood monocyte samples than with samples from *ex vivo* microglia, the implication is that the open

chromatin landscape of the iPSC_MGs more closely resembles that of cells derived from an earlier stage of the myeloid cell differentiation process (i.e. monocytes or macrophage precursor cells) than *ex vivo* microglia. This may reflect a limitation of the culture media in that the balance of factors within it may be insufficient to simulate the environmental conditions of the adult brain. Indeed, the fact that samples from both iPSC-derived lines cluster more closely with the *in vitro* microglia (which were shown to downregulate their microglial specific genes when placed in culture; 280), than the *ex vivo* microglia on PC2, supports this.

Microglial phenotype is shaped by the local microenvironment which includes physical and chemical interactions with neighbouring brain cells. Given the exquisite sensitivity of microglia to signals from other cell types, the fact that the iPSCs were not generated in co-cultures containing other cells could also explain why the open chromatin landscape of the iPSC_MGs does not recapitulate that of the *ex vivo* microglia. For example, neurons have been shown to maintain the adult homeostatic phenotype, and have an immunosuppressant effect on microglia through direct interaction between neuronal membrane bound factor CD200 and its target receptor CD200R expressed by microglia (290). Interleukin-34, a cytokine produced predominantly by neurons in the brain (267) also interacts with microglia to promote microglial differentiation and homeostasis (267,291).

Astrocytes also influence microglial phenotype via the activity of interleukins. Interleukin 33, which is released by synapse-associating astrocytes during post-natal neurodevelopment, has been shown to induce a synaptic engulfment phenotype in microglia via the IL1RL1 receptor in the thalamus (292). Conversely, interleukin-10 that is released by activated microglia during neuroinflammation, stimulates astrocytes to release TGF-β which, in turn, acts to reduce microglial activation and homeostatically resolve the neuroinflammatory process (293). Moreover, it has been reported that in order for microglia to sustain the typical ramified morphology associated with their homeostatic phenotype, they must be simultaneously exposed to TGF-β, GM-CSF and M-CSF which are all released by astrocytes in the brain (294). Intriguingly, in the study from which the iPSC protocol used here was partly derived (268), a co-culture differentiation strategy was adopted and iPSC_MGs were differentiated in the presence of neurons. In their principal component analysis comparing gene expression profiles of various iPSC-derived and primary cell types, Haenseler and colleagues demonstrated that iPSC_MG cells co-cultured with

neurons were distinct from iPSC_MΩpre cells and a iPSC_MG incubated in monoculture of iPSC_MΩpres with microglia medium (268). Therefore, although it is not possible to make a direct comparison between studies as different functional data were used to assess cell similarity, the data presented by Haenseler and collegues suggests that a co-culturing methodology may recapitulate the brain microenvironment better than using an iPSC monoculture strategy.

Considering the PCA results for the SV40 cells, all three samples cluster far from both the *ex vivo* microglia and the two iPSC lines suggesting that this line recapitulates the open chromatin landscape of *ex vivo* adult microglia the least out of all three cell lines. The lack of enrichment (and potentially expression) of the critical microglial lineage determining transcription factor Spi1 in the SV40 line (see table 3.5 and figure 3.1) suggests that the SV40 line has an altered gene regulatory profile when compared to the *ex vivo* microglia and iPSC-derived lines. This may be due to artefacts introduced during the immortalisation process and/or dedifferentiation that may have occurred in this cell line during culture (270,273). Indeed, in a recent study by Melief and colleagues comparing the gene expression profile of primary microglia to a suite of microglial cell culture lines, genes commonly identified as microglial-specific, and used as indicators for microglial specificity (*CX_3CR1, P2Ry12, MERTK, GAS6, PROS1 and GPR34*), were shown to be expressed at extremely low levels, or be undetectable, in SV40 cells when compared to primary microglia (275). As such, the results presented here are consistent with previous work.

There were several limitations of this study. First, only one biological replicate was used for each of the three cell culture lines generated. For the iPSC lines, this was due to the lack of cell availability, which was outside my control; however, in the case of the SV40 cells, I decided not to commit any further time or resources to this line once it was established that the motif analysis failed to find enrichment of Spi1. Having used only one biological replicate for each cell line, it is not possible to rule out that the findings reported here are due to specific batch effects introduced during cell culturing or the ATAC-seq laboratory protocol. A second limitation is that due to the differing iPSC-differentiation protocols used between this study and comparative microglia iPSC studies, any differences or similarities reported between the studies may be due to artefacts introduced using different methods. Third, as an iPSC's phenotype resembles that of a cell from an early stage in its developmental life cycle, the comparison between iPSC_MG cells and *ex vivo* microglia, which have been extracted from adults, may not be as appropriate as comparing IPSC_MG to primary

foetal microglia. Indeed, several iPSC microglial studies use foetal microglia as the main *ex vivo* dataset for comparison and report transcriptomic similarities between iPSC_MG and primary foetal microglia (268,278,295). However, one study reports a transcriptomic similarity between iPSC_MG and both adult and foetal primary microglia, suggesting that aspects of adult primary microglia phenotype can be recapitulated in iPSC_MG cells (277).

3.4.1 Future Work

Due to the highly dynamic nature of microglia, and the strong influence that other brain cell types have on the microglial phenotype, future work should consider generating iPSC-derived microglial lines in co-culture with multiple cell types, such as using a triple culture approach involving astrocytes and neurons. As yet, no triple-culture iPSC-derived microglia protocol has been published which is perhaps due to the fact that the factors required to differentiate non-microglial cell types can also compromise microglial function (e.g. Hanseler et al. omitted key neuronal supplement B27 from neuronal media for this reason; 263).

One alternative *in vitro* model is the cerebral organoid, which is a multicellular, three-dimensional structure resembling a cross-section of the embryonic human brain (296). Like iPSCs, cerebral organoids are generated from stem cells that, through the precisely timed application of Matrigel (components of extracellular matrix) and specific growth factors, are pushed toward a CNS fate. Organoids are considered a more physiologically relevant model than traditional 2D cell culture as their self-organisational properties promote spatial cell-cell interactions that more accurately recapitulate the *in vivo* environment. As such, organoids have been shown to contain mature neurons (and astrocytes) with functional synapses after 6 months growth (297,298). Regarding microglia, until recently, the consensus view was that organoid tissue could only generate cell lines that derived from ectodermal tissue, and as microglia derive from the mesoderm, organoids were not suitable for the study of microglial function. However, a recent study by Ormel and colleagues produced a method to generate mesodermal tissue within human organoids such that microglia were shown to emerge from mesodermal progenitors, develop ramifications and express critical factors including PU.1, IRF8 and RUNX1 and cell surface markers such as TREM2, CX_3CR1 and CD11b (299). Moreover, microglial processes were shown to interact with functional synapses in organoids, and synaptic debris was

detected in microglial vesicles implying that organoid microglia modulate and prune synapses in a similar fashion to *in vivo* microglia.

Whilst organoids do not entirely recapitulate the environmental conditions of the human brain, culturing microglia in 3D models provides a means to circumvent the critical issue in 2D models where microglia removed from the brain downregulate their microglial-specific genes when placed in culture conditions (157,265,280), as organoids are capable of intrinsically producing the factors that interact with microglia cell surface receptors to maintain their identify (i.e. CSF1 and Il-34). Several studies have emerged in recent years using organoids to model complex brain disorders. For example, Mariani et al. generated telencephalic organoids derived from individuals with ASD, who also displayed a macrocephaly phenotype, and healthy controls (300). Using a combination of cellular, transcriptome and RNA interference analysis Mariani et al., demonstrated that that there was an overproduction of GABAergic neurons and progenitor cells in organoids from ASD individuals and that this was driven by increased expression of the transcription factor FOXG1, supporting the hypothesis that an imbalance between excitatory and inhibitory neuronal activation states contribute to ASD aetiology (301).

3.4.2 Closing remarks

By comparing the open chromatin landscape of iPSC-derived microglial models, to that of *ex vivo* microglia and *in vitro* cells of the brain and circulatory system, I have established that the open chromatin landscape of iPSC_MG cells generated in monoculture (and following the Haensler et al. protocol), is more similar to that of cells antecedent to macrophages in the myeloid differentiation process (i.e. macrophage precursor cells or peripheral monocytes) than *ex vivo* adult microglia. I also demonstrate that the open chromatin landscape of SV40 cells does not recapitulate that of adult *ex vivo* microglia in line with previous work. In the next chapter, I will consider the open chromatin landscape of cryopreserved microglia extracted from the human foetal brain.

4 The open chromatin landscape of FACS sorted cryopreserved foetal microglia

4.1 Introduction

Producing novel functional genomic data from primary human brain tissue, and integrating it with robust genetic associations, will be essential to enhance our understanding of how non-coding genetic variation contributes to the onset and progression of complex brain disorders. Of particular importance will be determining the developmental and cellular contexts in which risk variants are active.

A prominent theory regarding the aetiologies of psychiatric disorders, particularly schizophrenia, ASD and ADHD, is that at least some of the perturbations that drive the onset of symptoms in later life occur during prenatal brain development (302,303). Evidence from epidemiological and animal studies have shown that maternal environmental insults, such as immune activation, during this period can increase the risk for schizophrenia and autism in offspring (230–234), and has led to the suggestion that synergistic effects between environmental stimuli and genetic susceptibility variants during this critical period can lead to the expression of disease-relevant endophenotypes (304). Microglial colonisation of the foetal brain occurs before neurogenesis and neural migration levels have reached their peak (305), which has led to the hypothesis that microglia have a pivotal role in shaping neuronal development (178,306). As described in chapter 1, microglia control key neurodevelopmental processes such as synaptic pruning and apoptosis (178,188). Furthermore, evidence across many complex brain disorders implicates the immune system as a key contributor to their genetic liability. As microglia are the predominant immune cell in the brain, it is plausible that non-coding risk variants that operate in microglial regulatory regions during this critical period of neurodevelopment could directly impact risk for a disorder, or synergistically operate via the environment to do so.

Several groups have recently integrated brain disorder GWAS data with functional information derived from human foetal brain tissue to shed light on early neurodevelopmental risk mechanisms for these conditions. For example, O'Brien and associates, mapped expression quantitative trait loci (eQTL) operating in the 2nd

trimester human brain, finding these to be enriched for common genetic risk variants for ADHD, bipolar disorder, and schizophrenia (307). Furthermore, de la Torre Ubieta and colleagues used ATAC-seq to measure chromatin accessibility in different regions of the mid-gestation human cerebral cortex, reporting that common variants associated with schizophrenia and ADHD were enriched in open chromatin sites in the germinal zone (associated with neural progenitor cells) but not enriched in those from the cortical plate (containing more mature neuronal cells; 292). However, a major drawback with these studies is that they could not determine how specific cell types contribute to the risk burden as their analyses were carried out on mixed populations of cells from bulk tissue. As gene regulatory processes such as chromatin accessibility, transcription factor expression and posttranslational modifications can be cell-type specific, genetic risk variants in non-coding regions could be regulated differently in distinct cell types. It is therefore of crucial importance to map cell-specific gene regulation in the developing human brain.

4.1.1 Aims

There is a current lack of functional genomic data derived from cells of the human foetal brain. Based on the hypothesis that foetal microglia mediate some of the genetic risk for brain disorders, the aim of this chapter was to produce a novel dataset that captured the open chromatin landscape of 'microglia' extracted from second trimester foetal human brain tissue for integration with genetic data. To do this, I extracted $CD11b^{+}$ cells from cryopreserved foetal brain cell suspensions using fluorescence activated cell sorting, and then mapped regions of open chromatin in extracted cells using ATAC-seq. In a similar manner to the partitioned heritability analysis in chapter 1, I tested foetal microglial open chromatin regions for enrichment of risk SNPs from a panel of brain disorder GWAS to ascertain whether risk for these disorders might be conferred through altered gene regulation in foetal microglia.

4.2 Methods

4.2.1 Samples

Human foetal brain tissue from elective terminations of pregnancy was provided by the Human Developmental Biology Resource (HDBR) (http://www.hdbr.org). The age (in post-conception weeks (PCW)) and sex of the 3 samples used in this study are shown in table 4.1 (information provided by the HDBR).

Table 4.1. Human foetal sample information

Sample ID	PCW	Sex	Hemisphere	Date received
14404	19	M	Right	17/10/2018
14473	17	M	Left	21/11/2018
14611	15	F	Right	06/02/2019

Sample ID (Sample identification number); PCW (post conception weeks); Sex (sex of foetus); Hemisphere (Brain hemisphere received for analysis).

4.2.2 Foetal tissue dissociation

Human foetal brain samples arrived as a single intact hemisphere that was immediately weighed (range 12.98-18.45g) and the meninges carefully removed. To homogenise the tissue, a scalpel was used for gross dissection and the dissected tissue was added, in batches if required, to a dounce homogeniser until the volume of tissue reached approximately 1cm below the fill line. Hibernate-E (Gibco) was added to the douncer until the fill line of was reached, and then 6 strokes of pestle A were applied to the douncer, followed by 6 strokes using pestle B. The homogenate was then transferred to a 50ml Falcon tube. This process was repeated until all tissue was homogenised. The total volume of the homogenate was then adjusted to ~45ml by adding Hibernate-E. For storage, 1ml aliquots of the tissue homogenate, containing 6% DMSO (Sigma), were taken and transferred into cryovials (Griener). Each vial was inverted 5 times to mix and all vials were cryopreserved in a Nalgene Mr Frosty container containing isopropanol which gradually cools the homogenate at a rate of -1^{o}C/minute. Cryovials were kept at -80^{o}C until required for further processing. Prior to cryopreservation, approximate cell counts were taken for each sample and these ranged from ~35-45 million cells per ml.

4.2.3 Fluorescence activated cell sorting (FACS)

Fluorescence activated cell sorting (FACS) is a flow cytometry method that allows the separation of individual cell populations from a heterogenous mixture of cells based on cellular light scatter and fluorescence characteristics. During this process, mixed cells passing through the sorter are hydrodynamically focussed to shuttle past a laser beam in a steady stream, one cell at a time. As cells are moved through the sorter's interrogation point, light from the laser is transmitted onto each cell and the scattering patterns of the deflected photons are used to measure the size and structural complexity of the cells (309). Cells can be further distinguished based on fluorescence characteristics. To facilitate this, antibodies that recognise a target feature on the outer membrane of a cell of interest are pre-labelled with fluorescent compounds called fluorophores, and then incubated with the mixed cell population. As the antibody should only bind to the cell of interest, when mixed cells pass the laser, electrons in the fluorophore are excited into higher orbitals by high energy photons in the laser light. Given that electrons in high orbitals are unstable, the electron quickly returns to its ground state and emits lower energy light, or fluoresces, as it does so. The light that the fluorophore emits can be used to distinguish fluorescently labelled cells from non-labelled cells. A range of fluorophores are available with different excitation and emission spectra which makes it possible to sort multiple cells types during one FACS experiment.

In preparation for FACS, 10-15 Eppendorf tubes were filled and incubated overnight (under rotation) with FACS buffer (0.5% bovine serum albumin, 2mM EDTA and PBS, pH 7.2). The following morning, 3 cryovials of dissociated cryopreserved foetal cells were placed in a 37^{o}C water bath and rapidly defrosted. The approximate cell number at the start each assay was 105-135 million cells. To accord with the recommended antibody dilution and maximum cell input recommended by the antibody supplier (1:50 for up to 10 million cells), defrosted cells were split evenly between 11-14 pre-coated Eppendorf tubes and centrifuged at 3200rpm for 5 mins at 4^{o}C. Identical centrifuge parameters were used throughout this protocol. The supernatant was removed from each tube and the cell pellets retained. Cells were then resuspended and washed in FACS buffer and centrifuged. After the supernatant was removed, cell pellets were resuspended in FACS buffer and incubated with antibodies targeting the myeloid cell surface receptor CD11b, for 10 minutes. The CD11b antibodies were pre-conjugated to the fluorophore APC (CD11b-APC; Miltenyi Biotech 130-110-612) for detection by FACS. The final

dilution of antibody in FACS buffer was 1:50. Cells were again washed in FACS buffer, centrifuged and the supernatant was removed. All cells were then pooled in 3ml of FACS buffer and placed in a pre-chilled FACS tube.

Cells were sorted on a FACS Aria III cell sorter using 100µm nozzle with the tube temperature set to 4°C. No compensation was required as only one fluorophore was used during sorting. Cells were sorted at a rate of ~20 million per hour for a maximum of 3 hours.

4.2.4 FACS gating

The gating strategy that was used for sorting $CD11b^{+}$ cells is depicted in figure 4.1 and was a 3-stage process.

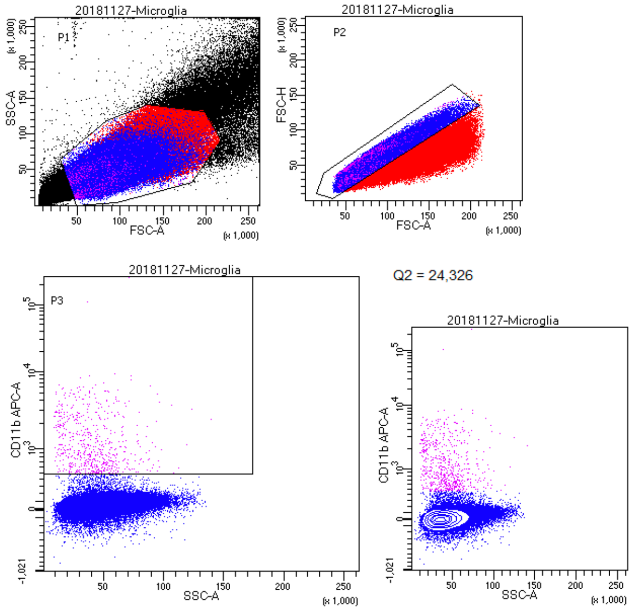

Figure 4.1: Representative gating strategy for sorting cryopreserved foetal cells by CD11b-APC fluorescence. Top left side scatter area (SSC-A) against forward scatter area (FSC-A). Top right forward scatter height (FSC-A) against forward scatter area. Bottom left and right APC fluorescence (CD11bAPC-A) against side scatter area. Black dots either cellular debris or cell doublets. Red dots cell doublets. Blue dots, individual foetal brain cells with low APC fluorescence and deemed CD11b-. Pink dots individual foetal brain cells with height APC fluorescence and deemed CD11b+. Inclusion gates shown in black outline.

Firstly, cells were sorted on forward scatter area (FSC-A) and side scatter area (SSC-A) (see figure 4.1 P1); the former estimates the size of each cell and the latter the structural complexity of the cells (i.e. their granularity). Due to the high cell input level it is not possible to see clustering of distinct cell types based on these parameters. However, this gate was used primarily to omit cellular debris, indicated in black and located at the bottom left hand side of the plot, and cell doublets, also indicated in black but located to the right, and upper right-hand side of plot A. To further enhance the doublet exclusion criteria the area scaling parameter of the sorter was used, which when set, produces a forward scatter height (FSC-H) and FSC-A correlation value for cells of the same size. Singlet cells would be located in a diagonal smear between 45^{o} and 90^{o}. In figure 4.1 P2, cells in blue were considered singlets and included in sorting, whereas red cells were considered doublets, or clumps of cells, and excluded from the sort.

The final gate was set based on the CD11b-APC fluorescence level. Two distinct clusters of cells were seen when CD11b-APC fluorescence level was plotted against SSC-A (see figure 4.1 P3): a large dense cluster with low APC fluorescence ($CD11b^{-}$ cells), indicated in blue, and a small sparsely distributed cluster with relatively high APC fluorescence ($CD11b^{+}$ cells), indicated in pink. Although the threshold distinguishing $CD11b^{+}$ from $CD11b^{-}$ cells was arbitrarily set, it was set conservatively to exclude as much of the large $CD11b^{-}$ population as possible.

FACS sorted, $CD11b^{+}$ cells were collected in FACS buffer, and once the sorting process was complete, all $CD11b^{+}$ cells were centrifuged, the supernatant removed, and the cells immediately used for ATAC-seq library preparation.

4.2.5 ATAC-seq library preparation

ATAC-seq library preparation for the FACS sorted $CD11b^{+}$ cells followed the OMNI-ATAC protocol (282), exactly as described in 3.2.4.

4.2.6 Sequencing, QC and bed file preparation

Paired-end 75 base sequencing was carried out on all 6 foetal brain derived microglial ATAC-seq libraries using the Illumina HiSeq4000. Sequencing files were de-

multiplexed, and then sequencing adapters and poor-quality reads were removed using Trim Galore (www.bioinformatics.babraham.ac.uk/projects/trim_galore/). In total, 12 fastq files were generated, each containing either the forward or reverse reads for a particular ATAC-seq library. The sequencing depth of each file was between 70-114 million reads. For quality control, FastQC (237) was run separately on all fastq files and MultiQC (238) used to collate the FastQC output. All files passed the quality control measures.

After quality control, read pairs for all samples were matched and aligned to the human genome (hg19; average alignment score >80%), using Bowtie2 (239) and mitochondrial reads were removed using Samtools (240). For peak calling, MACS2 (241) was run using `BAMPE` to handle paired-end reads and the FDR set to < 0.05 as the threshold for peak significance. Duplicate reads were ignored by MACS2 during the analysis by default. The final output from the peak calling process is a peak file.

In total, 6 unique peak files were obtained, corresponding to two technical replicate files for each of the 3 biological donors. Using these, I carried out a series of steps to obtain a single bed file that contained high confidence open chromatin regions operating in foetal microglia ($CD11b^+$ cells) that are not shared with other cells of the foetal brain. Initially, to improve the signal to noise ratio, technical replicate bed files for each donor were merged using Bedtools (310). Diffbind was then used to create a single consensus bed file by retaining peaks that were present in at least 2 of the 3 merged donor files (242). Next, to select open chromatin regions that are not shared with other cells of the developing human brain, the consensus file was intersected with an ATAC-seq peak file generated using bulk human foetal brain tissue (provided by my colleague Dr. Manuela Kouakou) and overlapping regions were removed (`bedtools intersect -v`). Finally, in order to retain only high confidence open chromatin regions within this 'microglia-specific' peak set, the file was intersected (`bedtools intersect -wa`) with a file containing genomic regions that are evolutionarily conserved in mammals, which was obtained from the 29 mammals project repository (https://www.broadinstitute.org/mammals-models/29-mammals-project-supplementary-info; 300). This final set of 35,466 peaks was used for all further analyses and will henceforth be referred to as 'conserved foetal microglia open chromatin regions'. Peaks were visualised as described in section 3.2.5.

4.2.7 Functional enrichment analysis

The online tool Genomic Regions of Enrichment Annotations (GREAT; 301) was used to test whether genes in proximity to conserved foetal microglia open chromatin regions are enriched within particular biological categories. GREAT functions by assigning all genes a basal regulatory domain that consists of a proximal domain (5kb upstream and 1kb downstream of the gene TSS) and a distal domain (1Mb upstream or downstream of the TSS) and calculates the enrichment of regulatory regions (provided by the user) within these domains using a binomial test that corrects for variability in gene regulatory domain size. GREAT was accessed (http://great.stanford.edu/) and run on May 23rd 2019, using the conserved foetal microglia open chromatin regions file and the whole human genome (hg19) as background.

4.2.8 *De novo* motif enrichment analysis

De novo motif enrichment analysis was used to test enrichment of transcription factor binding motifs within conserved foetal microglia open chromatin peak set. The HOMER package was used for this analysis as described in chapter 2.2.3.

4.2.9 Testing enrichment of risk variants for brain disorders within conserved foetal microglia open chromatin regions

SNPs within conserved foetal microglia open chromatin regions constituted 0.24% of the total SNPs contained within the 1000 genomes reference panel. As this proportion of SNPs within the functional annotation is below the recommended minimum of 1% for stratified LD score regression (sLDSR), an alternative method to test risk SNP enrichment within these sites was sought.

GARFIELD (GWAS analysis of regulatory or functional information enrichment with LD correction; 313)) is a software package designed to test for enrichment of GWAS risk variants in functional categories. Whereas sLDSR tests the enrichment of SNP heritability within functional annotations (i.e. the proportion of genetic variance in a trait explained by all SNPs), GARFIELD tests for enrichment of SNPs associated with the trait at specific p-value thresholds, by comparing these to an

independent set of SNPs matched for minor allele frequency, distance to the nearest transcription start site, number of linkage disequilibrium proxies and GC content.

To identify SNPs that overlapped regulatory regions for the GARFIELD analyses, custom annotation files were created. SNP information was provided with the GARFIELD software and was derived from the UK10K sequence data (314). Bedtools (310) was used to intersect SNP base positions with open chromatin regions derived from the ATAC-seq experiments, and each SNP was annotated with a 1 or 0 to indicate whether or not it fell within an open chromatin region. GWAS summary statistics were prepared using the `garfield-create-input-gwas.sh` script provided to retain only the base position and trait-associated p-value information for each SNP. GARFIELD required four additional files, all provided with the software, for the analyses. These included two LD tag files, derived from the UK10K study, that contain LD information for each SNP at two different thresholds (r^2 >= 0.1 and r^2 >= 0.8). To reduce the set of GWAS variants included in the analysis to a more computationally tractable independent set of SNPs, the lower LD r^2 threshold was used for a process called greedy pruning. This involved partitioning the genome into 1Mb windows, and for each window, sequentially removing SNPs with r^2 value > 0.1 relative to the most significant trait-associated variant and retaining the next most significant independent SNP (with r^2 < 0.1 being taken as an approximate measure of SNP independence) until a pruned set of independent SNPs was obtained. The higher threshold was used as the inclusion criteria for SNPs overlapping a functional region, i.e. a SNP was considered overlapping if it, or one of its LD proxies within 500kb (r^2 >= 0.8), was located within the open chromatin region. Finally, files to match variants by minor allele frequency and distance to the nearest transcription start site were included, the latter to correct for bias driven by local gene density. GARFIELD was applied to 7 GWAS of brain traits described in section 2.3.2. A GWAS of wearer of glasses or contact lenses (WGL; 249) which includes a similar number of genome-wide significant variants, but which does not appear to involve the immune system, was also tested as a negative control. To allow testing of traits with relatively low numbers of identified genome-wide significant SNPs, I tested enrichment of variants associated with traits at two P-value thresholds: $P < 1 \times 10^{-5}$ and $P < 1 \times 10^{-8}$. Statistical significance was calculated using a generalised linear model and enrichment relative to matched SNPs reported as an odds ratio. Empirical P-values

that survive Bonferroni-correction for 16 tests (8 traits at two GWAS P-values = P < 0.00313) are highlighted.

4.2.10 Overlap of open chromatin regions in conserved foetal and conserved adult microglia

To statistically test the overlap of open chromatin regions in conserved foetal and conserved adult microglia open chromatin regions, I performed a two-tailed Fisher's exact test using bedtools (310). The `bedtools fisher` command was run using the conserved foetal and adult microglial files, and the chromosome sizes of the genome the peaks were called from was also specified (hg19). Chromosome sizes are required to complete the Fisher's exact test contingency table and estimate the total number of possible intervals from which the peak intervals contained within each file could have derived. This makes it possible to estimate the number of intervals across the genome that are contained in neither peak file, which is required to complete the test contingency table.

4.3 Results

4.3.1 Characterisation of foetal microglial-specific ATAC-Seq peaks

Whilst the quality control procedures and data cleaning steps described in the methods (section 1.2.8) were designed to extract artefacts in the data (particularly contamination from other, non-microglia, brain cell types), it was important to further test that the FACS process was successful in isolating microglia. The online tool GREAT was therefore used to test whether genes that are proximal to identified open chromatin regions are enriched within particular biological categories of relevance to immune cells.

Figure 4.2. shows GREAT biological process terms that were most significantly associated with the foetal CD11b^{+} cell open chromatin regions. Five of the 12 most significantly enriched terms describe immune functions, including the regulation of myeloid leukocyte mediated immunity and macrophage differentiation. As these terms are immune cell-specific, and no non-immune brain cell-specific biological processes were reported (i.e. neuronal or oligodendrocyte specific processes),

these results suggest that the FACS process was successful in purifying $CD11b^+$ cells from the mixed foetal brain cell population.

Table 4.2. GREAT functional enrichment analysis of human foetal microglial open chromatin regions

Term Name	Enrichment	FDR cor. P-value
Preganglion parasympathetic nervous system development	3.18	2.37 x 10-41
Parasympathetic nervous system development	2.75	1.41 x 10-36
Macrophage differentiation	3.6	1.67 x 10-36
Filipodium assembly	2.72	6.31 x 10-32
Glossopharyngeal nerve development	3.6	2.77 x 10-27
Myoblast fate commitment	3.41	5.46 x 10-27
Response to type I interferon	2.26	7.71 x 10-27
Regulation of apoptotic process involved in morphogenesis	4.06	1.18 x 10-25
Glomerular endothelium development	6.68	2.56 x 10-25
Type I interferon-mediated signalling pathway	2.22	4.07 x 10-25
Cellular respone to type I interferon	2.22	4.43 x 10-25
Regulation of myeloid leukocyte mediated immunity	2.65	3.84 x 10-23

FDR cor. P-value (false discovery rate corrected p-value)

Considering the *de novo* motif enrichment results shown in table 4.3, foetal microglia open chromatin regions were significantly enriched for the myeloid fate determining factor Spi1 (target = 2.59%, background = 1.43%; p-value = 1×10^{-305}; (146,315). Similar to the GREAT analysis results in table 4.2, no transcription factor motifs specific to non-myeloid cells were identified in the motif enrichment analysis, providing further evidence that the $CD11b^+$ cell purification processing was successful.

Table 4.3. Motif analysis of human foetal microglia OCRs

Rank	Motif	% of Targets	% of Background	p-value
1	BORIS	2.97	0.95	1 x 10-1082
2	Zfp161	5.64	2.61	1 x 10-1061
3	Egr1	7.93	4.44	1 x 10-921
4	EGR1	4.77	2.38	1 x 10-747
5	HINFP	8.84	5.85	1 x 10-552
6	Spi1	2.59	1.43	1 x 10-305
7	ZNF189	3.86	2.73	1 x 10-166
8	Rfxdc2	16.46	14.32	1 x 10-140
9	Zfp691	13.83	12.04	1 x 10-115
10	NKX2	1.73	1.13	1 x 10-110
11	NPC	0.74	0.38	1 x 10-106
12	HOXA2	10.13	8.79	1 x 10-85
13	ZNF528	0.04	0.00	1 x 10-82
14	FXR	0.05	0.00	1 x 10-81
15	p53	4.10	3.30	1 x 10-73
16	Smad3	18.22	16.63	1 x 10-70
17	NFIX	0.03	0.00	1 x 10-66
18	bHLH	0.03	0.00	1 x 10-63
19	FOS::JUN	0.03	0.00	1 x 10-63
20	STAT1::STAT2	0.04	0.00	1 x 10-60

% of Targets (proportion of microglial open chromatin sites containing specified motif); % of Background (proportion of randomly selected background regions containg specified motif); OCR (open chromtin region)

When inspecting the foetal sample peaks tracks (see figure 4.2), similar to adult *ex vivo* and iPSC-derived microglia, there is a clear peak over the *SPI1* promoter indicating that myeloid specific open chromatin regions can be distinguished in the foetal samples.

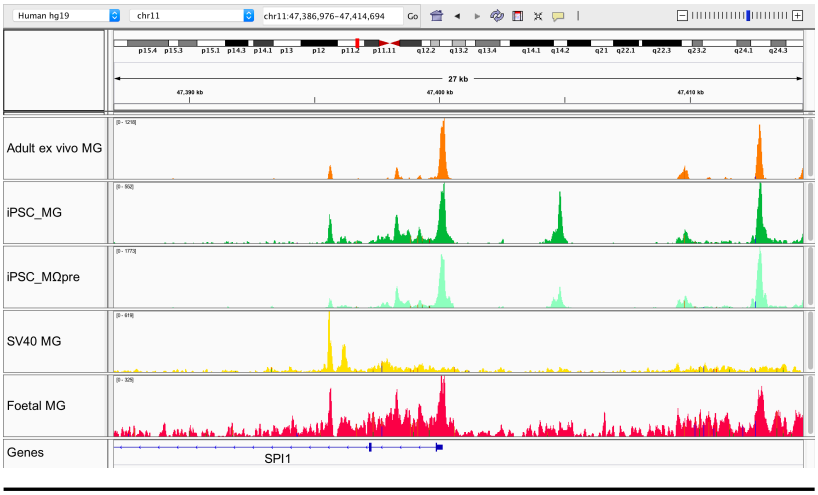

Figure 4.2. ATAC-seq peaks in adult *ex* vivo, foetal, and cell lines of, human microglia. Representative genomic tracks showing 27kb surrounding the promoter of the myeloid lineage determining transcription factor *SPI1*. X-axis, genomic location; Y-axis, auto-scaled read counts. Adult ex vivo MG (Adult *ex vivo* microglia); iPSC_MG (induced pluripotent stem cell derived microglia); iPSC_MΩpre (induced pluripotent stem cell derived macrophage precursor); SV40 MG (Simian virus 40 immortalised microglia); Foetal MG (cryopreserved 2nd trimester foetal microglia).

4.3.2 Principal component analysis

To compare the open chromatin landscape of the foetal microglia to that of the *ex vivo* and *in vitro* cell lines generated in chapter 3 (see section 3.3.2), the principal component analysis described in section 3.2.7 was performed again with the foetal sample data added.

The plot for the first two principal components (PCs) is shown in figure 4.3. As in the previous analysis (see figure 3.2), the first two principal components explained 85% of the total variance in the data. However, when the foetal data was added to the analysis PC1 and PC2 captured 78% and 7% of the variance respectively, as opposed to 77% and 8% when the foetal data was excluded. As PC2 separates cells of myeloid and lymphoid origin at 0.0, the fact that the foetal microglia samples cluster with myeloid cells provides additional evidence that the cell purification process was successful. Interestingly, on PC2 the foetal samples cluster more with the iPSC lines and the SV40 cells than the samples generated from adult microglia (*Ex-vivo* MG and *In-vitro* MG). This highlights that the open chromatin landscapes of foetal and adult microglia differ and suggests that open chromatin sites in human-

derived *in vitro* cell models are a better approximation of foetal microglia than adult microglia.

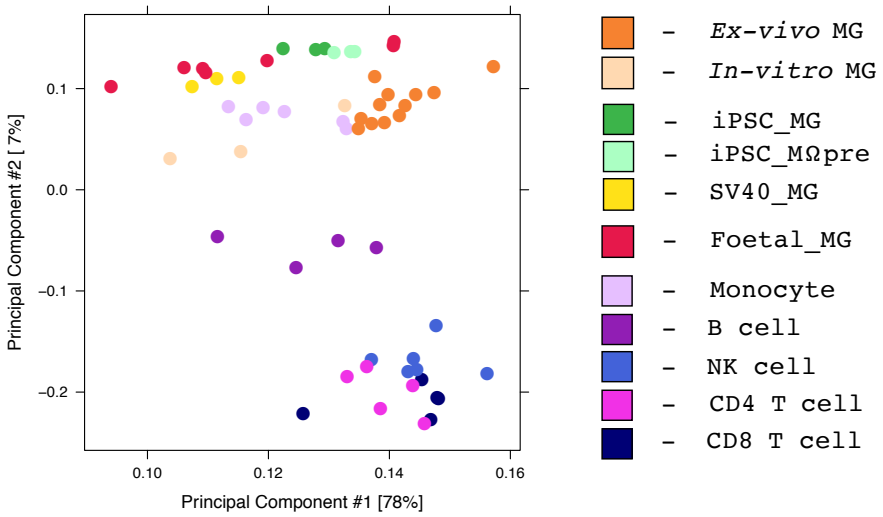

Figure 4.3. Principal component analysis plot of open chromatin site read counts derived from 11 cell lines. *Ex-vivo* MG (*ex-vivo* microglia); *In-vitro* MG (*in-vitro* microglia); iPSC_MG (induced pluripotent stem cell microglia; iPSC_MΩpre (induced pluripotent stem cell macrophage precursors); SV40_MG (SV40 immortalised microglia; Foetal_MG (foetal microglia); Monocyte (blood monocytes); B cell (B cells); NK cell (Natural Killer cells);. X-axis = principal component 1; y-axis = principal component 2. Read counts are $\log_2$ normalised.

Considering PC1, the foetal microglial samples are the most widely spread of all cell types in the analysis. The large variability between the open chromatin landscapes of the foetal samples may be explained by the fact that samples derived from donors of differing age and sex, or that the tissue for each donor derived from different brain hemispheres (see table 4.1), which may have affected microglial chromatin accessibility. Other factors such as the post-mortem delay period between the extraction of the foetal brain and its arrival for processing (usually 1-2 days) may also have contributed to this variability. Given that chromatin integrity and post-translational modifications have been reported to be stable for >48 hours post-mortem (316,317) it is unlikely that chromatin degradation underlies this variability. However, changes to the brain microenvironment during brain removal or transit may have induced chromatin accessibility changes in microglia, further contributing to between-sample variability.

4.3.3 Enrichment of bipolar disorder and schizophrenia risk SNPs in conserved foetal microglia open chromatin regions

GARFIELD was used to test for enrichment of SNPs taken from 7 brain disorder GWASs, and a negative control (described in chapter 2.3.2), at 2 GWAS significance thresholds, in conserved foetal microglia open chromatin regions. Table 4.4 shows the GARFIELD output for the 16 SNP enrichment tests carried out.

Table 4.4 Enrichment of brain disorder GWAS SNPs in conserved foetal microglia open chromatin regions

GWAS	GWAS Thresh	OR	p-value	cor. p-value	Beta	SE	CI95_L	CI95_U	No. Annot Thesh	N. Annot	N. Thresh	No. of SNPs
ADHD	1 x 10-5	1.62	0.352	1.000	0.483	0.519	-0.535	1.500	4	2,443	138	233,701
ADHD	1 x 10-8	7.41 x 10-8	0.998	1.000	-16.418	6173	-12116	12083	0	2,443	4	233,701
AUTISM	1 x 10-5	0.80	0.822	1.000	-0.229	1.016	-2.219	1.762	1	2,996	103	357,689
AUTISM	1 x 10-8	7.10 x 10-7	0.997	1.000	-14.158	4327	-8494	8466	0	2,996	1	357,689
BPD	1 x 10-5	2.97	**8.97 x 10-5**	**1.43 x 10-3**	1.087	0.278	0.543	1.631	15	8,461	227	1,191,616
BPD	1 x 10-8	7.76	**0.017**	0.268	2.049	0.856	0.371	3.728	2	8,461	8	1,191,616
LOAD	1 x 10-5	2.23	0.063	1.000	0.804	0.433	-0.045	1.652	6	2,874	113	262,885
LOAD	1 x 10-8	5.06 x10-8	0.994	1.000	-16.800	2380	-4680	4647	0	2,874	43	262,885
MDD	1 x 10-5	2.08	0.087	1.000	0.732	0.428	-0.107	1.570	7	8,343	159	1,157,454
MDD	1 x 10-8	6.70	**0.022**	0.357	1.902	0.832	0.271	3.534	2	8,343	8	1,157,454
NEUROTICISM	1 x 10-5	1.30	0.252	1.000	0.259	0.226	-0.184	0.702	24	8,063	697	907,858
NEUROTICISM	1 x 10-8	1.55	0.402	1.000	0.436	0.520	-0.583	1.455	5	8,063	84	907,858
SCZ	1 x 10-5	1.67	**0.003**	**0.045**	0.511	0.171	0.176	0.847	39	2,895	1,042	317,182
SCZ	1 x 10-8	1.68	0.117	1.000	0.520	0.332	-0.131	1.171	10	2,895	232	317,182
WGL	1 x 10-5	1.78	0.570	1.000	0.580	1.022	-1.423	2.583	1	5,389	68	891,538
WGL	1 x 10-8	8.60 x 10-7	0.997	1.000	-13.966	3762	-7387	7359	0	5,389	2	891,538

GWAS Thresh (GWAS significance threshold); OR (Odds ratio); p-value (emprical p-value of the significance of the observed enrichment); cor. p-value (p-value Bonferroni corrected for 16 tests); Beta (effect size from the general linear model, equal to log(OR)); SE (standard error from generalised linear model); CI95_L (lower boundary for the 95% CI of the effect size/beta); CI95_U (upper boundary for the 95% CI of the effect size/beta); No. Annot. Thresh (number of annotated variants within annotation passing GWAS thresh); N.Annot (number of variants within given annotation); N.Thresh (number of variants passing GWAS thresh); No. of SNPs (total number of LD pruned variants). ADHD (Attention deficit hyperactvity disorder); ASD (Autism spectrum disorder); BPD (Bipolar disorder); LOAD (Late onset Alzheimer's disorder); MDD (Major depressive disorder); SCZ (Schizophrenia); WGL (Wearer of glasses or contact lenses).

SNPs associated with bipolar disorder at both the $P < 1 X 10^{-8}$ and $P < 1 X 10^{-5}$ GWAS P-value threshold were significantly enriched in conserved foetal microglia open chromatin regions ($P < 0.05$), with the significance of the latter surviving Bonferroni correction for 16 tests (odds ratio = 2.97; $P = 8.97 X 10^{-5}$). SNPs associated with schizophrenia at the $P < 1 X 10^{-5}$ GWAS P-value threshold were also enriched at a P-value that survived correction for multiple testing (odds ratio = 1.67; P = 0.003). While significant enrichment of SNPs associated with major depressive disorder at the $P < 1 X 10^{-5}$ GWAS P-value (P = 0.022) was also observed, this did not survive correction for multiple testing. However, it is noted that the MDD and BPD SNP enrichment test statistics for SNPs at the $P < 1 X 10^{-8}$ threshold may have reduced reliability given that only 2 GWAS-associated SNPs fell within conserved foetal microglial open chromatin regions for these tests. Conserved foetal microglial open chromatin regions were not enriched for SNPs associated with any other trait, including the negative control of wearer of glasses or contact lenses (WGL).

4.3.4 Evolutionary conservation does not account for bipolar disorder / schizophrenia SNP enrichment signals in foetal microglial open chromatin regions

As previous studies have reported that GWAS SNPs for a variety of human traits are enriched in genomic regions that are highly conserved in mammals (244), it was important to test whether the SNP enrichment signals for bipolar disorder and schizophrenia reported in section 4.3.3 were driven by open chromatin regions in microglia rather than conserved sequence alone. To test this, fastq files for ATAC-seq data derived from 5 human adult non-brain tissues (oesophagus, colon, heart, liver, stomach) were obtained from the ENCODE repository (https://www.encodeproject.org). These files were processed in exactly the same way as described in section 4.2.9 for the foetal ATAC-seq data, including the foetal bulk extraction and evolutionary conserved region intersection steps. GARFIELD was run on the on all tissues and GWASs using the same parameters outlined in section 4.2.9.

As before, enrichment of SNPs associated with the 8 traits were tested at two GWAS significance thresholds: $P < 1 X 10^{-8}$ and $P < 1 X 10^{-5}$. In contrast to the enrichment observed within conserved foetal microglia open chromatin regions, SNPs associated with bipolar disorder were not enriched in conserved open chromatin regions in any of the 5 ENCODE tissues at even nominal significance (all $P > 0.05$). SNPs associated with schizophrenia were enriched in conserved open chromatin regions from some, but not all, ENCODE tissues, and for colon the significance survived Bonferroni correction for 16 tests (GWAS $P < 1 \times 10^{-5}$, OR = 1.90, $P = 2.97 \times 10^{-4}$; GWAS $P < 1 \times 10^{-8}$, OR = 3.18, $P = 5.80 \times 10^{-5}$; see chapter 8.1, tables 8.1-8.3). LOAD associated SNPs were also enriched in open chromatin sites in 2 of the 5 ENCODE tissues tested, with the test statistic in the liver surviving Bonferroni correction (GWAS $P < 1 \times 10^{-5}$, OR = 6.26; GWAS $P < 1 \times 10^{-5}$, OR = 12.24; see chapter 8.1, tables 8.4-8.5); however as only 3 or less LOAD associated SNPs were captured in conserved adult non-brain tissue OCRs the validity of these test results will have to be independently verified. GWAS-associated SNPs for the other brain traits were not significantly enriched in conserved open chromatin regions from any ENCODE tissue after multiple testing correction, (see chapter 8.1, tables 8.1-8.5), implying that the inclusion of evolutionary conserved regions is not

sufficient to drive SNP enrichment signal in open chromatin regions from non-relevant tissues.

4.3.5 Enrichment of brain disorder risk SNPs in conserved open chromatin regions from adult microglia

As it has been widely suggested that microglia have distinct functions in the foetus and in the adult (306), I tested whether the bipolar disorder and schizophrenia GWAS SNP enrichment seen in the foetal CD11b^{+} cell regulatory regions were also present in conserved open chromatin regions derived from an adult microglia dataset.

The adult *ex vivo* microglia ATAC-seq dataset was generated by Gosselin and associates as described in chapter 2.2.1 (236). The data was processed, and GARFIELD was run, exactly as described in chapter 4.3.4 with the exception that ATAC-seq regions from an adult bulk brain dataset, rather than the foetal bulk brain dataset, were used to exclude open chromatin regions shared with other neural cells (of the same developmental stage). This adult brain ATAC-seq dataset was obtained in peak file format from the CommonMind Consortium repository (https://www.synapse.org/#!Synapse:syn18134202).

Table 4.5 shows the GARFIELD results for the enrichment tests with the panel of brain disorder GWAS SNPs and the conserved adult *ex vivo* microglia open chromatin regions. Conserved adult microglia open chromatin sites are enriched for SNPs associated with several brain traits. For example, as for conserved foetal open chromatin sites, conserved adult microglia regulatory regions were also enriched for bipolar disorder (OR = 3.74, p = 7.21 x 10^{-5}) and schizophrenia (OR = 2.54, p = 6.91 x 10^{-7}) GWAS SNPs at the 1 x 10^{-5} threshold. However, by contrast, enrichment of bipolar disorder SNPs also were seen at the 1 x10^{-8} threshold (OR = 15.55, p = 1.27 x 10^{-3}) in adult microglia; this was not seen in the foetal CD11b^{+} cell analysis.

Table 4.5 Enrichment of brain disorder GWAS SNPs in conserved adult microglia open chromatin regions

GWAS	GWAS Thresh	OR	p-value	cor. p-value	Beta	SE	CI95_L	CI95_U	No. Annot Thesh	N. Annot	N. Thresh	No. of SNPs
ADHD	1 x 10-5	1.48	0.587	1.000	0.392	0.721	-1.021	1.805	3	1,270	138	233,701
ADHD	1 x 10-8	7.60 x 10-8	0.999	1.000	-16.392	8767	-17198	17165	0	1,270	4	233,701
AUTISM	1 x 10-5	7.71	**1.14 x 10-4**	**1.82 x 10-3**	2.042	0.529	1.005	3.080	4	1,457	103	357,689
AUTISM	1 x 10-8	1.13 x 10-6	0.998	1.000	-13.697	6288	-12338	12310	0	1,457	1	357,689
BPD	1 x 10-5	3.74	**7.21 x 10-5**	**1.15 x 10-3**	1.319	0.332	0.668	1.970	10	4,273	227	1,191,616
BPD	1 x 10-8	15.55	**1.27 x 10-3**	**0.020**	2.744	0.851	1.075	4.412	2	4,273	8	1,191,616
LOAD	1 x 10-5	3.06	**0.032**	0.507	1.118	0.520	0.098	2.138	4	1,568	113	262,885
LOAD	1 x 10-8	1.79	0.570	1.000	0.581	1.024	-1.425	2.588	1	1,568	43	262,885
MDD	1 x 10-5	4.74	**9.92 x 10-5**	**1.59 x 10-3**	1.556	0.400	0.772	2.339	7	4,240	159	1,157,454
MDD	1 x 10-8	7.98 x 10-8	0.997	1.000	-16.346	4499	-8833	8800	0	4,240	8	1,157,454
NEUROTICISM	1 x 10-5	1.28	0.421	1.000	0.248	0.308	-0.356	0.852	14	4,176	697	907,858
NEUROTICISM	1 x 10-8	2.24	0.174	1.000	0.808	0.594	-0.356	1.973	4	4,176	84	907,858
SCZ	1 x 10-5	2.54	**6.91 x 10-7**	**1.11 x 10-5**	0.931	0.188	0.563	1.299	33	1,544	1,042	317,182
SCZ	1 x 10-8	2.41	**0.017**	0.272	0.878	0.369	0.155	1.602	9	1,544	232	317,182
WGL	1 x 10-5	9.88 x 10-6	0.974	1.000	-11.525	356	-708	685	0	2,625	68	891,538
WGL	1 x 10-8	6.90 x 10-7	0.998	1.000	-14.186	5639	-11066	11038	0	2,625	2	891,538

GWAS Thresh (GWAS significance threshold); OR (Odds ratio); p-value (emprical p-value of the significance of the observed enrichment); cor. p-value (p-value Bonferroni corrected for 16 tests); Beta (effect size from the general linear model, equal to log(OR)); SE (standard error from generalised linear model); CI95_L (lower boundary for the 95% CI of the effect size/beta); CI95_U (upper boundary for the 95% CI of the effect size/beta); No. Annot. Thresh (number of annotated variants within annotation passing GWAS thresh); N.Annot (number of variants within given annotation); N.Thresh (number of variants passing GWAS thresh); No. of SNPs (total number of LD pruned variants). ADHD (Attention deficit hyperactvity disorder); ASD (Autism spectrum disorder); BPD (Bipolar disorder); LOAD (Late onset Alzheimer's disorder); MDD (Major depressive disorder); SCZ (Schizophrenia); WGL (Wearer of glasses or contact lenses).

Furthermore, unique to the adult *ex vivo* microglia analysis, enrichment of autism (OR = 7.71, p = 1.14 x 10^{-4}) and major depressive disorder (OR = 4.80, p = 9.92 x 10^{-5}) associated SNPs were seen in conserved open chromatin regions at the P < 1 x 10^{-5} threshold. Consistent with the stratified linkage disequilibrium score results reported in chapter 2.3.2, SNPs associated with Alzheimer's disease at the P < 1 X 10^{-5} threshold were also enriched in open chromatin regions from adult microglia (P = 0.032), although the significance of this observation did not survive correction for multiple testing. However, the reliability of the autism, Alzheimer's disease and major depressive disorder enrichment results (as well as the BPD result for SNPs at the P < 1 x 10^{-8} threshold) could be questioned due to the low number of SNPs located in the conserved foetal open chromatin regions annotation.

These results suggest that some of the SNPs that are most strongly associated with bipolar disorder and schizophrenia impact gene regulatory processes in both foetal and adult microglia, and that gene regulation in adult microglia may also be affected by variants associated with autism and major depressive disorder.

4.3.6 Overlap of open chromatin regions in conserved foetal and conserved adult microglia

Given that the GARFIELD analyses indicated that bipolar disorder- and schizophrenia- associated SNPs were enriched in both conserved foetal and conserved adult *ex vivo* microglial open chromatin regions, I decided to formally test the overlap of open chromatin regions of both cell types. Figure 4.4 shows that of

the 35,137 and the 13,301 peaks contained in the conserved foetal and conserved adult *ex vivo* microglia peak files respectively, 2,206 peaks overlapped both peak sets. This overlap is more than expected by chance given the genomic coverage of the peak intervals (P < 0.0001, Fishers Exact Test). However, there are also many peaks that are only identified in one cell type; these may reflect genuine physiological differences between foetal and adult microglia, differences in sample processing or experimental noise.

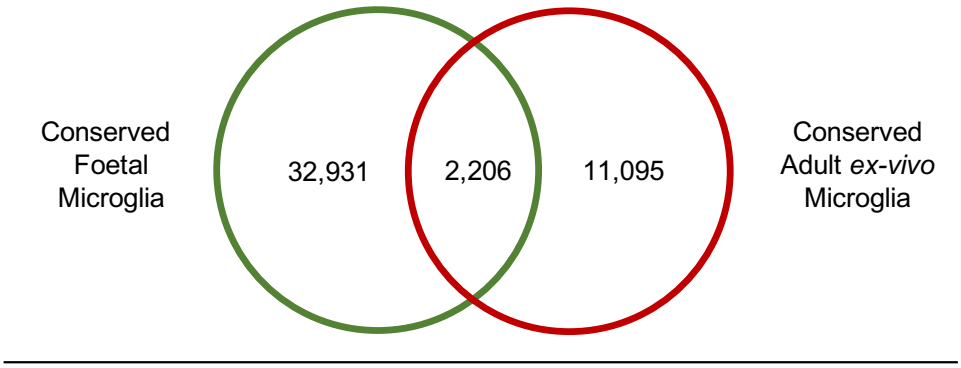

Figure 4.4. Overlap of open chromatin regions in foetal and adult ex-vivo microglia. Foetal microglia (conserved foetal microglia open chromatin regions with foetal bulk regions removed); Adult *ex-vivo* microglia (conserved adult *ex-vivo* microglia open chromatin regions with adult bulk regions removed).

4.4 Discussion

Establishing which brain cell types mediate genetic risk is an important step in improving our understanding of causal mechanisms for complex brain disorders. The evidence presented in section 4.3.2. suggests that a significant proportion of common genetic variants associated with schizophrenia and bipolar disorder plausibly operate through $CD11b^{+}$ cells - the vast majority of which are likely to be microglia (318,319) - during prenatal brain development.

Microglia display a broad spectrum of phenotypes throughout life and, as such, modulate their gene expression profile according to the stimuli they perceive in their local microenvironment. Transcriptomic studies in mice suggest that microglia alter their gene expression profile, and presumably open chromatin landscape, in a temporally sequential manner, expressing proliferative genes during the embryonic phase, genes involved in synaptic pruning during the foetal phase, before settling into the adult-like, homeostatic, gene expression profile soon after birth (141). Given

that the putative biological cause of many psychiatric disorders is aberrant neural wiring or connectivity (320,321), microglia's role during the foetal period in regulating neuronal and synaptic maturation is of particular interest as (A) it provides plausible mechanism through which a non-neuronal cell can contribute to neuron pathology and (B) suggests that timing of exposure to genetic and/or environmental insults may be critical for increasing liability for these disorders.

Evidence from epidemiological and animal studies support the hypothesis that peri-natal perturbation of microglia environment is a risk factor for psychiatric brain disorders. For example, environmental insults such as pre-natal maternal stress or immune activation (i.e. caused by infection) have been shown to increase the risk for psychiatric disorders such as schizophrenia, autism and ADHD (230–234) in offspring. Furthermore peri-natal maternal stresses have been shown to have long term effects on microglial morphology (322), gene expression profile, phagocytic activity (323,324), and perturb the maturation process of microglia in progeny (141,322). The data presented here provide additional, functional evidence, suggesting that genetic factors in microglial-specific open chromatin regions, which are likely to impact the regulation of gene expression, also contribute to disease risk during this critical period in neurodevelopment. Moreover, given that the open chromatin sites overlap between foetal CD11b^{+} cells and adult *ex vivo* microglia (reported in section 4.3.6), and the fact that GWAS risk SNP enrichment was seen in adult microglia open chromatin regions for several brain disorder traits, these data suggest that genetic risk factors for neuropsychiatric disorders may operate in microglia throughout life.

Although this study provides novel functional genomic information from human foetal immune cells, providing better cellular resolution than similar studies carried out using bulk brain tissue (307,308,325), it does not resolve distinct, or niche, populations of myeloid cells that exist in the brain. While microglia are by far the largest population of myeloid cells in the brain, additional myeloid cell populations expressing CD11b exist near blood vessels, the choroid plexus and subarachnoid space (326). Due to this, many groups use a dual antibody FACS strategy to isolate microglia by gating using CD11b and CD45, where low and high expression of CD45 is used to differentiate microglia and from other myeloid cells respectively (327). I did initially attempt to use CD45 in my gating strategy (see figure 4.5) but had to abandon this due to the extremely low cellular yields.

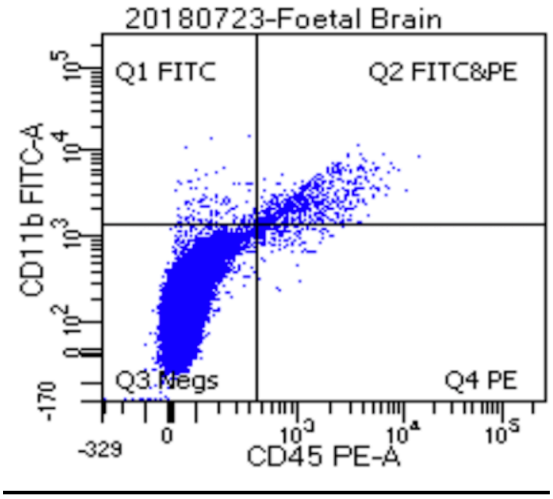

Figure 4.5. FACS image showing dual antibody gating strategy. X-axis = PE fluorescence indicative of CD45 cell surface expression level (CD45 PE-A); Y-axis = FITC fluorescence indicative of CD11b cell surface expression level (CD11b FITC-A). Q1, FITC positive cells; Q2, FITC and PC double positive cells; Q3, FITC and PC double negative cells; Q4, PE positive cells.

As I optimised the FACS protocol, it became clear that the proportion of foetal microglia in the human foetal brain (<0.1%) was considerably lower than that reported in adult human brain samples (5-10%). The lower proportion of foetal microglia compared to adult microglia has been corroborated recently in a single cell analysis (328) and should be considered in future attempts to isolate microglia from mixed foetal cell suspensions. Moreover, it has been suggested that distinct, brain region-specific, populations of microglia exist that differ in gene expression (329) and gene enhancer profiles (139,236,330). Indeed, a recent single cell transcriptomic study suggested that microglia heterogeneity may be most marked in the developing brain (331). These distinctions could not be resolved in this study as $CD11b^+$ cells were extracted from whole brain hemispheres.

Another limitation of the study is that the cell isolation methods used to isolate 'microglia' in the foetal brain tissue described here and in adult brain tissue described in the Gosselin et al. study were different (236). For example, Gosselin and colleagues added a cell enrichment step, based on cellular mass, using a viscosity gradient and used 4 cell surface markers ($CD11b^+$, $CD45^{low}$, $CD64^+$,

CX3CR1^{High}) for their FACS isolation. Given the issue, discussed above, regarding the relatively lower number of CD11b^{+} cells in the foetal brain compared with the adult, these additional steps were not possible in this analysis.

Unlike the sLDSC (see section 2.4), no basement threshold for SNPs coverage within a functional annotation of interest is stated in the GARFIELD documentation. Indeed, as GARFIELD tests statistical significance of SNP enrichment by running 10,000 permutations on randomly generated SNPs that have been matched in number and by feature (minor allele frequency, distance to the nearest gene and local LD), SNP coverage is accounted for in this test (332). Despite this, compared to sLDSC, GARFIELD is relatively new methodology and has not been yet been rigorously tested in the field. As such, the enrichment results reported here require independent validation.

4.4.1 Future Work

Following on from the work outlined above, the next logical step would be to attempt to characterise individual microglial cell populations in the 2nd trimester brain based on their transcriptomic, or epigenomic, profiles using single cell sequencing. This could be done by brain region, at different post-conception stages. Single cell approaches would be well suited for characterising foetal microglia as they require a relatively low cell input number. Furthermore, it is now possible to perform ATAC-Seq at the single cell level (333), enabling insight into gene regulatory mechanisms at play in distinct subpopulations of microglia, and their potential role in complex brain disorders.

4.4.2 Concluding remarks

Using ATAC-seq, I have generated novel epigenomic data derived from CD11b^{+} cells extracted from human 2nd trimester foetal brain. I found that GWAS risk SNPs associated with bipolar disorder and schizophrenia are enriched in conserved open chromatin sites within these cells, suggesting that foetal microglia mediate some of the genetic risk for these disorders.

5 Electromobility mobility shift assay (EMSA) on candidate risk variants for Alzheimer's disease

5.1 Introduction

As the vast majority of common psychiatric risk associated loci are located outside regions that encode genes, determining the functional impact any mutation has at a molecular level is challenging (334). This is exacerbated by the fact that most risk associated loci contain thousands of SNPs that are in linkage disequilibrium making it difficult to ascertain causal, from associated, SNPs. Unlike coding mutations which directly affect protein formation, conformation or enzymatic activity, non-coding variants are most likely to impact risk by altering the molecular processes that modulate gene expression (90).

A fundamental process that facilitates gene expression, and one of the key drivers of phenotypic variation, is the binding of transcription factors to regulatory elements in euchromatic regions of the genome (116). As discussed in chapter 1.3, at regulatory elements such as enhancers, DNA binding transcription factors interact with short sections of DNA sequence called motifs and, once bound to DNA, recruit other regulatory proteins to drive or repress gene expression. Through chromatin looping, distal regulatory elements are brought into close spatial proximity with the promoters of their target genes and the transcription factors bound to each regulatory element interact to alter gene expression. One mechanism by which non-coding SNPs could increase genetic susceptibility for a psychiatric disorder is by altering the consensus sequence of a transcription factor motif (335) . For example, considering a motif located within an enhancer, a single base pair change from A to G could alter the consensus sequence of the binding site such that transcription factor binding is more or less likely (or prevented altogether), with consequent effects on gene expression.

5.1.1 Aims

Having shown in chapter 1 that >50% of common variant LOAD heritability was contained in open chromatin sites of *ex vivo* microglia, the aim of this chapter was to attempt to characterise how LOAD associated variants impact microglial function at the molecular level. Based on the rationale that non-coding risk alleles are likely to alter the consensus sequence of transcription factor binding sites (90), a set of candidate LOAD risk alleles that fell within (A) an adult microglial open chromatin site and (B) a transcription factor motif relevant to microglial function, were identified. Computational analysis was then used to prioritise the alleles that were predicted to have the most deleterious impact on motif binding. By designing DNA oligonucleotides containing risk or alternate alleles at each prioritised locus, and incubating these oligonucleotides separately with microglial nuclear extracts, electrophoretic mobility shift assays (EMSAs) were carried out to assess DNA-protein interactions at each locus. This made it possible to assess whether nuclear protein binding is differentially affected in a cellular system carrying the LOAD risk allele compared with the alternate allele, and to potentially identify cell relevant transcription factors that may be impacted. Functional experiments such as this are important as they provide insight into the molecular basis of genetic risk, and also provide a means to prioritise the risk SNPs that are most relevant to disease risk in the post-GWAS era.

5.2 Materials and Methods

5.2.1 Prioritising LOAD risk SNPs located in microglial regulatory regions using MotifbreakR

To prioritise candidate SNPs located in microglial regulatory regions for functional analysis, a list of LOAD genome-wide significant index SNPs from the IGAP GWAS (80) and correlated SNPs in LD ($R^2 \geq 0.8$) were collated, numbering 390 SNPs. As the SNP location information was obtained in Chr:Start:Stop format, BiomaRt (336) was used to map SNP locations to rsIDs. Next, Bedtools intersect (310) was used to select the SNPs that fell within adult *ex vivo* microglia open chromatin regions. A description of how the chromatin accessibility data were processed can be found in Chapter 1 and in the Tansey et al. study (235). This identified 29 SNPs, from 11 LOAD

genome wide significant loci, that overlapped putative regulatory regions in adult microglia.

MotifBreakR is a bioinformatics tool that annotates any list of SNPs to known transcription factor motifs genome-wide (337). Moreover, if any SNP falls within a motif, MotifbreakR reports the name of the motif and a score describing how strongly each allelic variant modifies the binding potential of the factor/s that putatively interact at the motif. MotifbreakR was chosen for this task as it has been previously used for functional annotation of GWAS data (338), it is easier to automate, and it queries more transcription factor motif repositories than similar online annotation tools such as RSAT (339).

The MotifbreakR pipeline was a three-stage process. First, the rsIDs for the LOAD index SNPs overlapping microglial regulatory regions were annotated with location and allelic information using the dbSNP142 (340) and these locations used to query build hg19 of the human reference genome for sequence information surrounding the variant sites. Second, the sequence information identified for each locus was cross referenced with 4 transcription factor motif repositories (ENCODE (341), Factorbook (342), HOCOMOCO (343), Homer (121)) to identify motifs containing the risk SNPs that we wish to query for potential disruption of transcription factor binding. Finally, a position probability matrix was produced for each motif putatively disrupted, which models the relative affinity, and the probability of occupancy, of a transcription factor for a consensus sequence (344). By summing the probability scores for the most frequently occurring base at each position along the motif, MotifbreakR derives an optimal motif score for each motif. It then generates two further motif scores (i.e. it substitutes each allelic variant from the SNP of interest into the optimal motif), which makes it possible to compare each allele-specific motif score to the optimal motif score. This allows MotifbreakR to model the level of disruption caused by each allele at that motif. The allelic motif scores were scaled from 1-0, and any allele with a score of $\geq$0.85 was reported. A p-value threshold of 1×10^{-4} was set and 370 potentially disruptive motifs, each of which contained one of the original 29 GWAS SNPs, reported. For each of the 390 entries, the ratio of the motif scores for the reference and alternate alleles was taken, and the SNP/motif at each extreme was selected after sorting entries by score ratio. These SNPs were deemed the best candidates to take forward as their alleles had the largest differences between their motif scores and thus the most likely to have a detectable difference in the EMSA assay.

5.2.2 Validation of prioritised SNPs using data from public repositories

In order to corroborate whether the SNPs that had been prioritised by the MotifbreakR analysis were suitable for experimental follow up, two public repositories were accessed.

Haploreg v4.1 is a repository used for assigning functional annotations to non-coding genetic loci. By combining LD information from the 1000 genomes project with chromatin state and protein binding information generated by the Roadmap Epigenomics and ENCODE projects, Haploreg allows researchers to generate hypotheses about how non-coding variation impacts phenotype (345). Haploreg was used to corroborate the MotifbreakeR results for both rs2883470 and rs938152 (https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php accessed 01.04.19).

Cistrome is a public database containing over 20,000 ChIP-seq and chromatin accessibility datasets that can be exploited for gene regulatory analysis (346,347). As, at the time of writing, the only human microglia chromatin based study available was that produced by Gosselin and colleagues (236), Cistrome was used to access chromatin state data in human macrophages, a cell type of the same lineage. These data were accessed on 01.04.19 (http://cistrome.org/db/#/).

5.2.3 Nuclear protein extraction and quantification

To harvest enough BV2 nuclear protein for multiple EMSAs, three ~80-90% confluent T125 flasks of BV2 cells were harvested as described in chapter 2. Ice-cold NE-PER nuclear and cytoplasmic extraction reagents (Thermo Fisher – 78833) were used to extract the BV2 nuclear protein. BV2 cells were first centrifuged at 16,000g for 5 minutes to dry the cells as much as possible and transferred to a 1.6ml Eppendorf tube. Next, 1ml of CER I was added to the cell pellet. The pellet was vortexed on the highest setting for 15 seconds to ensure it was thoroughly resuspended and then it was incubated on ice for 10 minutes. Then, 55µl of CER II was added, the nuclei were vortexed for 15 seconds, and then centrifuged at 16,000g for 5 minutes. The supernatant, which was the BV2 cytoplasmic extract was discarded. The nuclear pellet was suspended in 250µl of NER, half the recommended volume, then vortexed

for 15 seconds every 10 minutes for a total of 40 minutes. The sample was then centrifuged at 16,000g for 10 minutes and the supernatant, which was the BV2 nuclear protein extract, was transferred to a clean Eppendorf. The protein extract was quantified using the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific) and the nuclear extract was aliquoted into 150µg batches and snap frozen at -80°C.

5.2.4 Designing and annealing oligonucleotides

For each selected SNP (rs9381562 and rs28834970), complementary 50 nucleotide base biotinylated DNA oligonucleotides were synthesised along with unlabelled competitor oligonucelotides incorporating either the reference or alternate allele. In total, 16 oligonucelotides were created, see table 5.1 (only 8 shown).

Table 5.1 Oligonucleotides designed for electrophoresis mobility shift assay

SNP	Motif	Allele	Oligonucleotide
rs9381562	ETS	C - Ref	AGTGTCAATGGTGGAACTGGGCTTC**C**TGGCTTCAGGAGTTCAGTCCAAGTG
			CACTTGGACTGAACTCCTGAAGCCA**G**GAAGCCCAGTTCCACCATTGACACT
		A - Alt	AGTGTCAATGGTGGAACTGGGCTTC**A**TGGCTTCAGGAGTTCAGTCCAAGTG
			CACTTGGACTGAACTCCTGAAGCCA**T**GAAGCCCAGTTCCACCATTGACACT
rs28834970	C/EBP	T - Ref	TGGTCATTCCATATAAGTGGAATTG**T**ACAACACTGTGGCCTTTCGCGACGG
			CCGTCGCGAAAGGCCACAGTGTTGT**A**CAATTCCACTTATATGGAATGACCA
		C - Alt	TGGTCATTCCATATAAGTGGAATTG**C**ACAACACTGTGGCCTTTCGCGACGG
			CCGTCGCGAAAGGCCACAGTGTTGT**G**CAATTCCACTTATATGGAATGACCA

DNA sequences for 50bp oligonucleotides designed for the LOAD associated SNPs rs9381562 and rs28834970. Reference and alternate alleles are shown in red. Oligonucleotides were either biotinylated or unlabelled at the 5' end (not shown).

For annealing, equimolar complimentary oligonucleotides were heated in a thermocycler to 95°C for 5 minutes then cooled 1°C every 5 minutes until they reached room temperature. Annealed oligonucleotides were diluted to a final concentration of 1pmol/µl in 18.2Ω H_2O.

5.2.5 Electrophoresis mobility shift assay

The EMSA detects DNA-protein interactions based on the fact that DNA-protein complexes migrate more slowly than unbound DNA in a polyacrylamide gel, resulting in a shift in migration of labelled, protein-bound, oligonucleotide. The LightShift® Chemiluminescent EMSA Kit (Thermo Fisher Scientific) was used for the DNA-protein binding reactions and visualisation steps of the EMSA procedure.

A 10 well, 5% Mini-PROTEAN® TBE precast DNA gel (Bio-Rad) was placed in an electrophoresis unit filled with 0.5X TBE. The 30µl wells were flushed thoroughly and 100V was passed through the gel for 30 minutes. Next, 5µl of 5X loading dye was added to each sample followed by gentle mixing by pipetting, the wells were flushed, and 20µl of each sample was loaded into separate wells of the gel. Again, 100V was passed through the gel until the bromophenol blue dye had migrated ¾ the length of the gel (approximately 45 minutes). Meanwhile a positively charged Biodyne B Nylon Membrane (Thermo Fisher Scientific) was soaked in 0.5X TBE for 10 minutes. Once migration was complete, the gel was floated off carefully from its plastic casing in a bath of 0.5X TBE. The gel was then placed on top of the nylon membrane, and blotting paper applied to each side. Bubbles were removed using a roller and the sandwich was placed into a mesh lined cassette. The cassette was placed into a clean electrophoresis unit filled with 0.5X TBE (chilled to ~10^{o}C) and 100V was applied for 30 minutes. When transfer had occurred, the excess TBE was removed from the membrane using a paper towel and the DNA was crosslinked to the membrane using a UV-light set to 120mJ/cm^{2} for 1 minute.

5.2.6 Chemiluminescence reaction and visualisation

Blocking and wash buffers were heated to ~37^{o}C until particulates had completely dissolved. The membrane was then passed through a series of incubation/wash steps all of which included gentle shaking on a plate shaker. These included a 15-minute incubation in blocking buffer followed by a further 15-minute incubation in stabilised streptavidin-horseradish peroxidase conjugate diluted 1:300 in blocking buffer. The membrane was then rinsed briefly in 1X wash buffer and placed in 1X wash buffer for five minutes. This five-minute wash was repeated three times. Next, the membrane was bathed in substrate equilibration buffer (SEB) for 5 minutes. In the final step, excess SEB was drained from the membrane before it was incubated, without shaking, with substrate working solution (Luminal/Enhancer Solution and Stable Peroxide Solution mixed at a 1:1 ratio) for five minutes.

The chemiluminescence on the membrane was detected and visualised using the Syngene G-Box Chemi XX9 system.

5.2.7 Quantification of DNA

Image J was used to quantify DNA captured during imaging. For standardisation, images were first transformed to the 16-bit format (Image > Type > 16-bit), then calibration measurements were taken from the least and most saturated regions respectively. A separate calibration curve was generated for each image (Analyse > Calibrate > straight line) and arbitrary units for intensity maxima and minima were set from 0 to 100. The rectangle tool was used to measure DNA quantity by capturing the relative intensity of each band. Two-sample t-tests were performed to compare band intensities between reference and alternate alleles for each SNP.

5.2.8 Optimisation of EMSA

To optimise the EMSA a total of eight 20µl samples were set up as shown in table 5.2 and figure 5.1. The first three reactions included a DNA only control, containing biotinylated DNA but no nuclear protein extract or unlabelled competitor DNA, a DNA-protein reaction containing biotinylated oligonucleotide and nuclear extract, and an unlabelled competitor reaction containing biotinylated oligonucelotide, nuclear extract and unlabelled target DNA. The latter acted as a competitive inhibitor to the biotinylated DNA. For comparison, an additional five DNA-protein reactions were set up, each containing a single optimisation reagent, to assess if one or more of these reagents improved the DNA-protein interaction in this system. As indicated in table 5.2, 4pmols of unlabelled target oligonucelotide, 10µg of nuclear protein extract and 20fmol of Biotin labelled oligonucelotide where added to the appropriate reactions. This optimisation step was carried out for risk and non-risk alleles for both SNPs (4 optimisation experiments in total).

Representative, DNA-protein interactions for the ALT allele at rs9381562 and ALT allele at rs28834970 are shown in figure 5.1, marked with arrows. These images demonstrate that the EMSA can effectively detect DNA-protein interactions at these

loci. In both cases, the interaction was reduced markedly by the addition of unlabelled competitor oligonucleotide (lane 3) indicating the specificity of the interaction at each locus. Regarding the 5 optimisation reagents, it was decided that the clearest banding was evident when NP40 was added (lane 5), particularly for reactions involving rs9381562. Henceforth, 1µl of NP40 was added to all reactions.

Table 5.2. Reagents for EMSA optimisation

Reagent	Amount
H_2O (18.2Ω)	Make up to 20µl
10X Binding Buffer	2µl
Poly (dI•dC) 1µg/µl	1µl
Glycerol (50%) *	1µl
NP-40 (1%) *	1µl
KCL (1M) *	1µl
$MgCl_2$ (100mM) *	1µl
EDTA (200mM) *	1µl
Unlabelled DNA (1pmol/µl)	4µl
BV2 nuclear protein extract	10µg
Biotin labelled DNA	20fmol

* optional

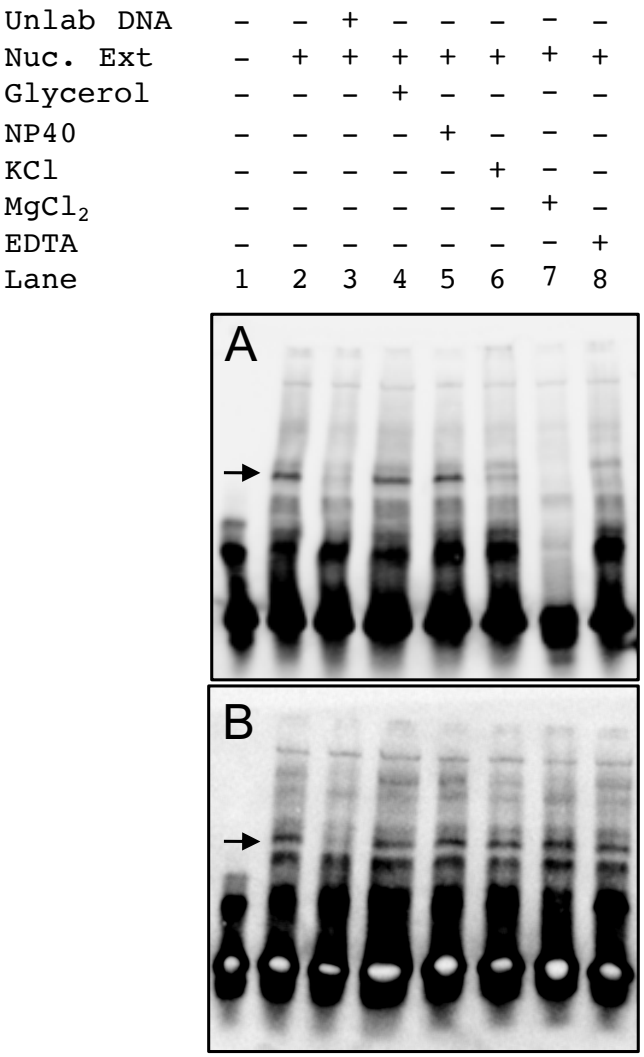

Figure 5.1. EMSA optimisation in BV2 cells. DNA-protein interactions involving oligonucleotides containing rs9381562 (A) and rs28834970 (B) ALT alleles. Oligonucleotides in all wells are dsDNA except those marked with * which are ssDNA.

The optimised EMSA was performed on reference and alternate alleles for both rs9381562 and rs28834970 and three technical replicates were obtained for each individual allele.

5.2.9 Supershift assay

In addition to the normal EMSA procedure, a series of supershift assays were included for oligos containing rs9381562 alleles. During the supershift assay, an antibody targeting a protein of interest is added to the EMSA reaction. If the target protein is present in the protein network bound to the biotinylated oligonucelotides, the complex would migrate more slowly through the membrane due to the additional mass of the bound antibody. This would lead to a supershift of the DNA-protein band in lanes containing the antibody compared to lanes containing control antibodies. The supershift assays were initially carried out following the EMSA procedure outlined above, however, after the 20-minute protein binding incubation, 2µg of either PU.1 antibody (Santa Cruz – sc352) or the control IgG antibody (inhouse rabbit polyclonal) were added to appropriate wells and incubated for a further 30 minutes. The water content of the affected wells was adjusted to keep the total volume constant across all wells. The remainder of the EMSA procedure was carried out as normal.

5.3 Results

5.3.1 Prioritisation of LOAD associated risk SNPs for molecular analysis using MotifbreakR

The motifbreakR analysis identified 390 motifs that were putatively disrupted by the 29 LOAD risk SNPs included in the analysis. To identify the SNPs that were the best candidates to take forward for further analysis, a ratio of the motif scores for each allele at each motif were taken. The rationale for this was that, as the motif score is a proxy for the binding potential of each putative motif, motifs with the largest difference in their allele-specific motif scores would represent the best candidates to take forward for EMSA as they would be the most likely candidates to have observable allelic differences in DNA-protein binding. Table 5.3. shows the MotifbreakR output for the SNPs rs9381562 on chromosome 8 and rs28834970 on chromosome 6. Motifs at the extremes of table 5.3 contain the largest differences in their allele-specific motif scores.

Table 5.3 Abridged MotifbreakR output showing putatively disrupted motifs identified for the LOAD associated risk SNPs rs28834970 and rs9381562.

rsID	Chr	REF	ALT	Motif	Sequence	Ref score	Alt score	Ratio	Effect
rs28834970	chr8	T	C	CEBPB	gaattgTacaacac	5.68	7.32	0.78	strong
				CEBPD	aattgTacaaca	5.75	7.39	0.78	strong
				CEBPA	attgTacaac	5.17	6.56	0.79	strong
				EP300	aattgTacaacactg	6.04	7.63	0.79	strong
				NA	agtggaattgTacaa	6.19	7.80	0.79	strong
				CEBPB	attgTacaac	7.57	9.38	0.81	strong
				DDIT3::CEBPA	ggaattgTacaa	6.91	8.38	0.82	strong
				DDIT3::CEBPA	tggaattgTacaa	6.93	8.40	0.82	strong
				NA	aattgTacaacact	7.71	9.35	0.82	strong
				CEBPE	ggaattgTacaa	6.58	7.88	0.84	strong
				CEBPG	attgTacaac	8.59	10.01	0.86	strong
				STAT3	attgTacaa	10.45	12.09	0.86	strong
				DBP	aattgTacaac	5.55	6.39	0.87	strong
				DDIT3	tggaattgTacaa	7.49	8.62	0.87	strong
				CEBPD	attgTacaaca	6.27	7.16	0.87	strong
				CEBPG	attgTacaac	8.45	9.63	0.88	strong
				CEBPG	ggaattgTacaac	7.26	8.21	0.88	strong
				CEBPB	attgTacaac	7.71	8.65	0.89	strong
				CEBPE	attgTacaac	7.13	7.85	0.91	strong
				BPTF	ttgTacaacactgt	4.51	4.80	0.94	weak
				CEBPA	gaattgTacaacac	4.56	4.84	0.94	weak
				CEBP	gaattgTacaacac	4.60	4.81	0.96	weak
				NFAT5	gtggaattgTacaa	12.13	11.13	1.09	strong
				ETS1	ggaattgT	5.07	4.59	1.10	weak
				ZNF143	tggaattgTa	11.42	9.78	1.17	strong
rs9381562	chr6	C	A	NR2F2	gggcttcCtggcttca	8.79	10.43	0.84	strong
				EGR1	actgggcttcCtggcttcaggagtt	15.72	15.40	1.02	weak
				NA	gcttcCtggcttcaggagttc	14.08	13.68	1.03	weak
				RAD21	cCtggcttcaggag	5.02	4.52	1.11	weak
				STAT3	ttcCtggct	10.97	9.86	1.11	strong
				ELF2	gaactgggcttcCtggct	12.62	10.99	1.15	strong
				ELF2	ctgggcttcCtggct	8.05	6.70	1.20	strong
				ELF3	gcttcCtggcttca	9.02	7.49	1.20	strong
				ERG	ggcttcCtgg	9.02	7.40	1.22	strong
				ETV4	gcttcCtg	6.08	4.99	1.22	strong
				ELK3	gggcttcCtggc	6.77	5.54	1.22	strong
				GRHL2	gaactgggcttcCtggcttc	7.69	6.22	1.24	strong
				ETS	gcttcCtg	6.49	5.18	1.25	strong
				ELF1	ggcttcCtgg	8.90	7.02	1.27	strong
				ETS	ctgggcttcCtg	7.43	5.79	1.28	strong
				ETS1	gcttcCtgg	7.68	5.75	1.34	strong

Chr (chromosome); REF (reference allele) ALT (alternate allele); Sequence (DNA sequence considered in analysis with SNP capitalised); Ref score (Ref motif score); Alt score (Alt motif score); Ratio (ratio of absolute Alt/Ref scores);

The variant rs9381562 (GWAS p-value = 7.83 x 10^{-8}) lies in a non-coding region of chromosome 6 and is in perfect LD (r^2 = 1) with the LOAD genome wide significant SNP rs10948363 (GWAS p-value 5.20 x 10^{-11}). At rs9381562, the most significantly disrupted motif that was proposed was the ETS1 motif which was contained within a 9bp sequence on the negative strand, see table 5.3. With a motif score of 5.75 (P = 5.34 x 10^{-5}), the A_{ALT} allele was more disruptive to the motif than the C_{REF} allele which scored 7.68 (P = 1.61 x 10^{-2}) representing a 25% reduction in binding potential. Notably, 3 of the top 4 most disrupted motifs identified at this SNP affect an ETS domain binding site.

Considering rs28834970, this variant falls within an intron of protein kinase 2 beta (PTK2β) and is a LOAD genome-wide significant SNP (GWAS p-value = 7×10^{-14}; 80). The motif highlighted as being the most significantly disrupted at this locus was the C/EBP-β domain located within a 14bp sequence on the negative strand (see table 5.2). The predicted direction of effect was opposite to that seen in rs9381562 with the T_{REF} allele being more detrimental to the binding potential of the motif than the C_{ALT} allele, with motif scores of 5.68 ($P = 7.65 \times 10^{-3}$) and 7.31 ($P = 6.17 \times 10^{-6}$) respectively. This corresponds to a 22% reduction in binding potential when the T_{REF} allele is substituted for the C_{ALT} allele. Interestingly, 15 of the 25 putatively disrupted motifs identified by MotifbreakR for rs28834970 impact motifs of the C/EBP family.

These results provide statistical evidence that rs9381562 and rs28834970 are suitable candidates to take forward for molecular analysis. Further biological evidence for their suitability from the literature will be discussed in subsequent sections.

5.3.2 Publicly available data suggests rs9381562 and rs28834970 fall within active regulatory regions in cells of a myeloid lineage

Given that a bioinformatics tool has been used thus far to prioritise SNPs for further analysis, it was important to find supporting evidence in the literature for the suitability of these SNPs for molecular analysis (348).

Chromatin accessibility and ChIP-seq data from the Glass et al. adult *ex vivo* microglia study (236) suggest that both rs9381562 and rs28834970 are located in enhancer regions in microglia. In figure 5.2 A, rs9381562 is situated within a chromatin accessibility peak (ATAC) implying that this SNP falls within an active, or poised, regulatory element in microglia. Moreover, as the rs9381562 is also located at the edge of peaks associated with the histone modifications H3K4Me2 and H3K27ac, it is inferred that that this regulatory element is an active enhancer.

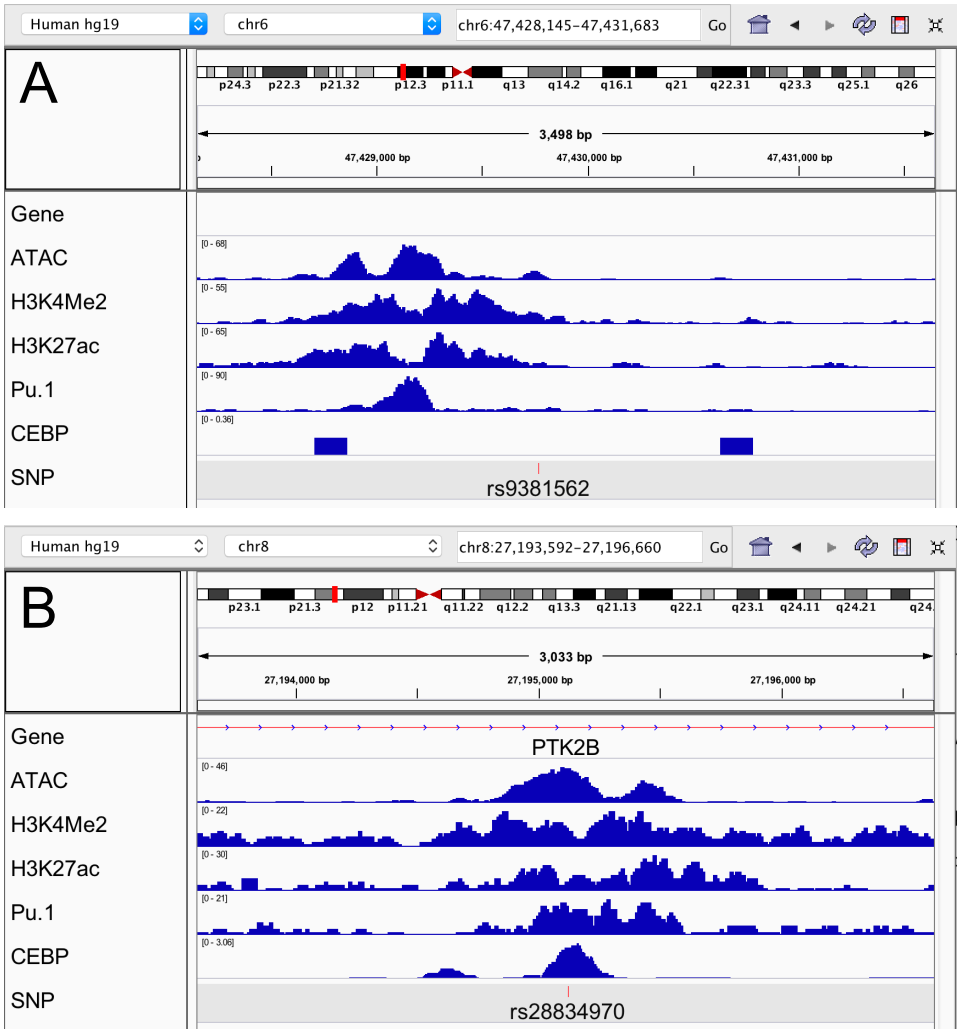

Figure 5.2. Chromatin accessibility and chip-seq data from publicly available sources. Representative genomic tracks showing approximately 3kb surrounding the LOAD associated SNPs rs9381562 (A) and rs28834970 (B). X-axis is genomic location and Y-axis auto-scaled read counts. Peak height corresponds to the frequency with which each annotation occurs at a particular locus. Chromatin accessibility data (ATAC) and chip-seq data (H3K4Me2, H327ac and Pu.1) from human ex vivo microglia taken from Glass et al. 5.22017 (236). Chip-seq data for CEBP taken (347) from a study in macrophages.

Regarding rs28834970, figure 5.2 B, a chromatin accessibility peak, as well as H3K4Me2 and H3K27ac associated ChIP-seq peaks derived from human *ex vivo* microglia surrounding the SNP, suggest that this SNP is located in an enhancer region with regulatory potential in microglia. Moreover, rs28834970 is surrounded by a peak derived from a C/EBP-β transcription factor ChIP-seq study in macrophages (349), suggesting that C/EBP-β binds at this locus in myeloid cells. Similarly, the Haploreg v4.1 database provides evidence that rs28834970 is located in an active regulatory region in cells of myeloid lineage as it is associated with the histone marks H3K4Me1, H3K4Me3 and H3K427ac in primary myeloid cells from peripheral blood

(see figure 5.3). The Haploreg v4.1 output also suggests that the binding potential of C/EPB-β motifs are disrupted by rs28834970 in an allele-specific manner, with the motif scores for the for the T_{REF} allele being consistently lower than the C_{ALT} allele for all 4 C/EBP-β entries. This provides independent evidence that the T_{REF} allele at rs28834970 is more disruptive to the binding potential of the C/EPB-β motif than the C_{ALT} allele.

These results suggest that both rs9381562 and rs28834970 are located in active enhancer regions in microglia and suitable candidates for molecular analysis.

chr	pos (hg19)	chr	pos (hg38)	Reference	Alternate	1000 Genomes Phase 1 Frequencies				Sequence constraint		dbSNP functional annotation
						AFR	AMR	ASN	EUR	by GERP	by SiPhy	
chr8	27195121	chr8	27337604	T	C	0.22	0.3	0.23	0.35	No	No	intronic

Closest annotated gene

Source	Distance	Direction	ID/Link	Common name	Description
GENCODE	NA	Within gene	ENSG00000120899.13	PTK2B	PTK2B protein tyrosine kinase 2 beta [Source:HGNC Symbol;Acc:9612]
RefSeq	NA	Within gene	NM_173174	PTK2B	PTK2B protein tyrosine kinase 2 beta [Source:HGNC Symbol;Acc:9612]

Epigenome ID (EID)	Group	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac
E062	Blood & T-cell	BLD.PER.MONUC.PC	Primary mononuclear cells from peripheral blood		4_PromD2	H3K4me1_Enh		
E034	Blood & T-cell	BLD.CD3.PPC	Primary T cells from peripheral blood		12_TxEnhW	H3K4me1_Enh		H3K27ac_Enh
E045	Blood & T-cell	BLD.CD4.CD25I.CD127.TMEMPC	Primary T cells effector/memory enriched from peripheral blood		12_TxEnhW	H3K4me1_Enh		
E033	Blood & T-cell	BLD.CD3.CPC	Primary T cells from cord blood	6_EnhG	12_TxEnhW	H3K4me1_Enh		
E044	Blood & T-cell	BLD.CD4.CD25.CD127M.TREGPC	Primary T regulatory cells from peripheral blood		12_TxEnhW	H3K4me1_Enh		
E043	Blood & T-cell	BLD.CD4.CD25M.TPC	Primary T helper cells from peripheral blood		10_TxEnh5	H3K4me1_Enh		
E039	Blood & T-cell	BLD.CD4.CD25M.CD45RA.NPC	Primary T helper naive cells from peripheral blood	7_Enh	12_TxEnhW	H3K4me1_Enh		H3K27ac_Enh
E041	Blood & T-cell	BLD.CD4.CD25M.IL17M.PL.TPC	Primary T helper cells PMA-I stimulated		10_TxEnh5	H3K4me1_Enh	H3K4me3_Pro	H3K27ac_Enh
E042	Blood & T-cell	BLD.CD4.CD25M.IL17P.PL.TPC	Primary T helper 17 cells PMA-I stimulated		12_TxEnhW	H3K4me1_Enh		
E040	Blood & T-cell	BLD.CD4.CD25M.CD45RO.MPC	Primary T helper memory cells from peripheral blood 1	7_Enh	12_TxEnhW	H3K4me1_Enh		
E037	Blood & T-cell	BLD.CD4.MPC	Primary T helper memory cells from peripheral blood 2	6_EnhG	12_TxEnhW	H3K4me1_Enh		
E048	Blood & T-cell	BLD.CD8.MPC	Primary T CD8+ memory cells from peripheral blood	7_Enh	12_TxEnhW	H3K4me1_Enh		
E038	Blood & T-cell	BLD.CD4.NPC	Primary T helper naive cells from peripheral blood	6_EnhG	10_TxEnh5	H3K4me1_Enh		
E047	Blood & T-cell	BLD.CD8.NPC	Primary T CD8+ naive cells from peripheral blood	7_Enh	10_TxEnh5	H3K4me1_Enh		
E029	HSC & B-cell	BLD.CD14.PC	Primary monocytes from peripheral blood	7_Enh	9_TxReg	H3K4me1_Enh	H3K4me3_Pro	H3K27ac_Enh
E031	HSC & B-cell	BLD.CD19.CPC	Primary B cells from cord blood	3_TxFlnk	9_TxReg	H3K4me1_Enh	H3K4me3_Pro	
E035	HSC & B-cell	BLD.CD34.PC	Primary hematopoietic stem cells	1_TssA	10_TxEnh5	H3K4me1_Enh	H3K4me3_Pro	
E051	HSC & B-cell	BLD.MOB.CD34.PC.M	Primary hematopoietic stem cells G-CSF-mobilized Male	6_EnhG	10_TxEnh5	H3K4me1_Enh		
E050	HSC & B-cell	BLD.MOB.CD34.PC.F	Primary hematopoietic stem cells G-CSF-mobilized Female		10_TxEnh5	H3K4me1_Enh		H3K27ac_Enh
E036	HSC & B-cell	BLD.CD34.CC	Primary hematopoietic stem cells short term culture		4_PromD2	H3K4me1_Enh		
E032	HSC & B-cell	BLD.CD19.PPC	Primary B cells from peripheral blood	6_EnhG	10_TxEnh5	H3K4me1_Enh		
E046	HSC & B-cell	BLD.CD56.PC	Primary Natural Killer cells from peripheral blood	6_EnhG	10_TxEnh5	H3K4me1_Enh		
E030	HSC & B-cell	BLD.CD15.PC	Primary neutrophils from peripheral blood	2_TssAFlnk	9_TxReg	H3K4me1_Enh	H3K4me3_Pro	

Regulatory motifs altered

Position Weight Matrix ID (Library from Kheradpour and Kellis, 2013)	Strand	Ref	Alt	Match on: Ref: TTTCTGGTCATTCCATATAAGTGGAATTGTACAACACTGTGGCCTTTCGCGACGGCTGC Alt: TTTCTGGTCATTCCATATAAGTGGAATTGCACAACACTGTGGCCTTTCGCGACGGCTGC
CEBPA_2	+	10.4	11.8	RTTGYRHMW
CEBPB_disc1	+	13.8	14.8	RTTGYRCAAY
CEBPB_known1	+	0.4	11.8	NTTDCHHMABHH
CEBPB_known5	+	5.3	11.7	DKVTTRCDHMAYHN
CEBPB_known6	+	0.2	12.2	MBMTTDCHHMAYHY
CEBPD	+	-0.3	11.7	VATTDCDYMMYY
Hsf_disc1	+	11.9	12.8	VTTRYRYAAS
STAT_disc4	+	4.4	16.4	ATTRCWCAA
p300_disc2	+	0.9	12.7	NRTTKCAHMABHHHH

Figure 5.3. Haploreg v4.1 output for region surrounding rs28834970 in blood and immune cell types. Functional chromatin annotations associated with rs28834970, providing evidence that histone modifications indicative of active regulatory elements are found in myeloid cells from peripheral blood (E029 – primary monocytes). 7_Enh, H3K4Me1_Enh indicate enhancers, whilst H3K4Me3_Pro is indicative of promoters and H3K27ac_Enh is indicative of regulatory element activation. Output also shows that binding potential of the C/EBP-β motif is consistently lower for T_{REF} allele compared with C_{ALT} allele.

5.3.3 EMSA indicates that DNA-protein interactions at rs9381562 and rs28834970 are impacted in an allele-specific manner in BV2 cells

Incubation of BV2 nuclear extracts with biotinylated oligonucleotides containing either allele of rs9381562 produced two distinct bands that had significant allelic differences in the intensity of bound DNA (see figure 5.4 A). Considering the upper band, indicated by the upper arrow, the strong band seen in well 5 is partially eliminated in well 6 by the inclusion of 200-fold excess unlabelled competitor oligonucleotides. This indicates that the DNA-protein interaction seen here is specific to this DNA sequence. When testing differences in band intensity between reactions containing the A_{ALT} and C_{REF} alleles at rs9381562 (lanes 2 and 5 respectively), oligonucleotides encompassing the LOAD associated A_{ALT} allele bound, on average, 50% less nuclear protein(s) than oligonucleotides containing the C_{REF} allele ($p < 6.2 \times 10^{-4}$, t-test, figure 5.4 B). By contrast, the lower band, indicated by the lower arrow, had a significant allelic difference in band intensity in the opposite direction to the upper band, with DNA containing the A_{ALT} allele (lane 2) binding 22% more nuclear protein than oligonucleotides incorporating the C_{REF} allele (lane 5; $p < 0.02$, t-test).

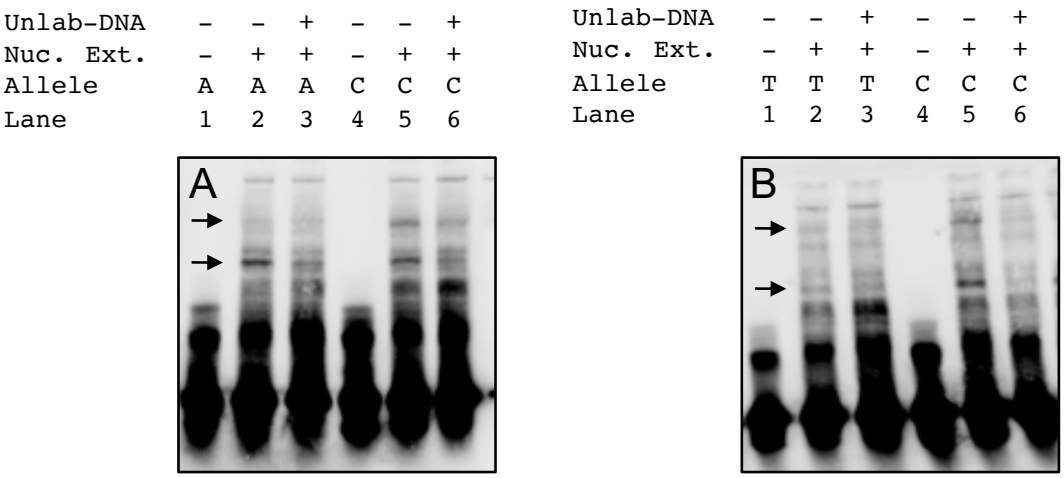

Figure 5.4. EMSA incubating oligonucleotides containing either rs9381562 (A) or rs28834970 (B) in BV2 cell nuclear extract. DNA-protein interactions involving oligonucleotides containing alleles from either rs9381562 (A) or rs28834970 (B) and nuclear factors from BV2 cells. All wells contain oligonucleotides that were biotin labelled. Unlab-DNA (Unlabelled DNA; added in 200-fold excess); Nuc. Ext. (Nuclear extract). DNA in all wells is dsDNA.

Similarly, two candidate bands, suggesting allele specific differences in DNA-protein binding, were seen in experiments involving rs22834790, see figure 5.4 B. The DNA-protein interaction depicted by the intensity of the lower band (lane 5, lower arrow) is abolished when excess unlabelled competitor oligonucleotide is added (lane 6, lower arrow), again indicating the sequence specificity of this interaction. Significant allelic differences were measured between the C_{ALT} and T_{REF} alleles when comparing band intensity (lanes 2 and 5 respectively), with the T_{REF} allele binding 58% less protein on average compared to the C allele ($p < 0.003$, t-test, figure 5.5 D).

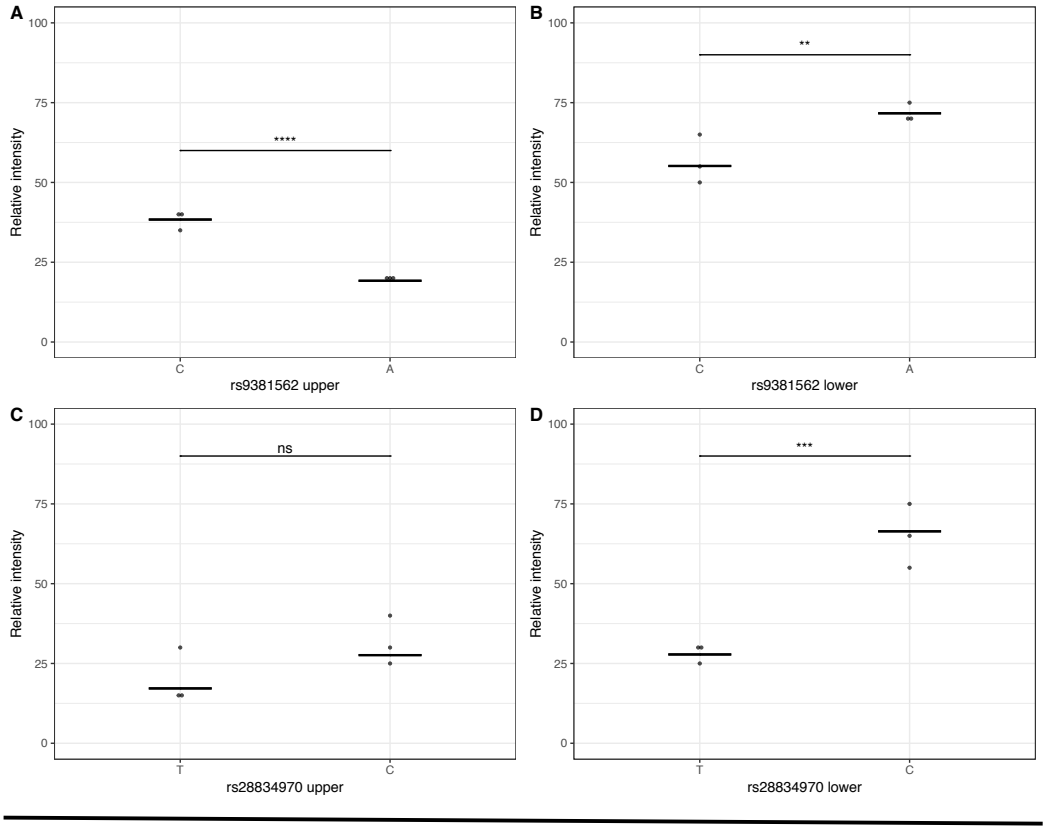

Figure 5.5. Quantification of binding of nuclear factors from BV2 cells to DNA oligonucleotides containing either rs9381562, or rs28834090, alleles. Y axis = relative intensity of DNA protein band, x axis = allele at each SNP locus. ** P < 0.01 , *** P < 0.001, **** P < 0.0001.

When testing allelic differences in DNA-protein binding in the upper band, although there was a trend in the same direction as seen in the lower band (i.e. DNA containing the C_{ALT} allele bound more protein than DNA incorporating the T_{REF} allele), this interaction was non-significant between the alleles (p = 0.17, t-test; figure 5.5 C).

These results suggest allele-specific DNA-protein interactions occur at both rs9381562 and rs28834970.

5.3.4 EMSA indicates that PU.1 may be present at rs9381562

Given the finding reported in chapter 1 that LOAD risk SNPs were localised in microglial open chromatin regions containing the transcription factor PU.1, and the fact that the MotifbreakR analysis highlighted that variation at rs9381562 putatively

impacted an ETS domain motif, of which PU.1 is known to bind to, I decided to carry out a EMSA supershift assay using the rs9381562 alleles.

Figure 5.6, shows a representative EMSA supershift reaction. Notably, when the antibody targeting PU.1 was added to reactions containing oligos with either the C_{REF} or A_{ALT} allele at rs9381562, lanes 7 and 9 respectively, no supershift banding pattern was seen, implying that PU.1 was not present within the DNA protein complex bound at this locus.

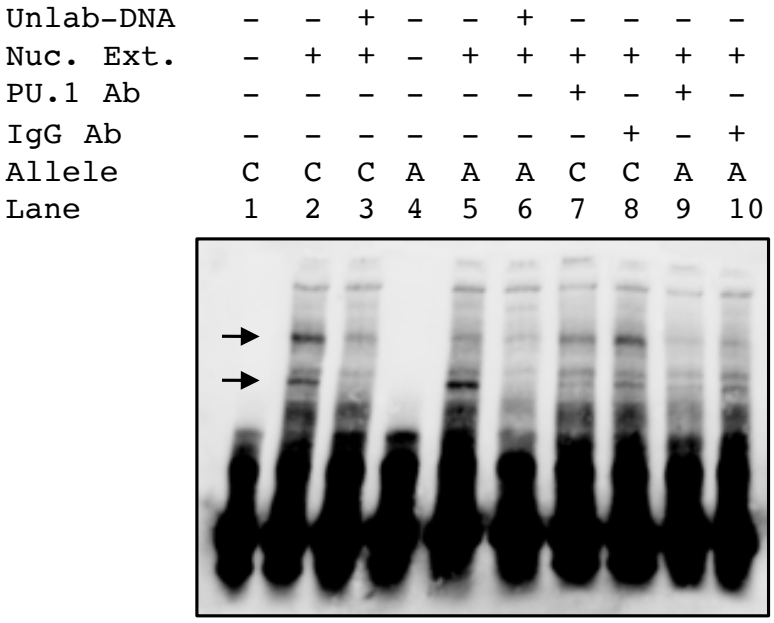

Figure 5.6. EMSA supershift assay incubating oligonucleotide containing rs9381562 alleles with nuclear factors from BV2 cells and either PU.1 or IgG antibodies. Allele-specific interactions indicated by arrows. DNA-protein interactions involving oligonucleotides containing rs9381562 and nuclear extract from BV2 cells. All wells contain oligonucleotides that were biotinylated. Unlab-DNA (Unlabelled DNA; added in 200-fold excess); Nuc. Ext. (Nuclear extract). DNA in all wells is dsDNA.

However, when considering the upper band (upper arrow) in figure 5.6, identical reactions containing C allele oligos incubated either with (lane 7) or without (lane 2)

the PU.1 antibody, indicate that band intensity is lessened when the PU.1 antibody is present. Moreover, when the IgG antibody is substituted for the antibody targeting PU.1, lane 8, band intensity recovers to a level similar to that seen in lane 2. This could suggest that PU.1 is present in the bound protein network, but that the antibody attaches to the protein in such a way that it prevents PU.1 interacting with either the DNA or the entire network.

5.4 Discussion

The aim of this chapter was to investigate how LOAD-associated polymorphisms impacted transcription factor binding in microglia. Establishing how individual cell types in the brain contribute to pathology is an important step in improving our understanding of the causal mechanisms of complex brain disorders. In particular, pinpointing the molecular events, within individual cell types, that are perturbed as these disorders progress is critical, both to the process of characterising reliable biomarkers for diagnosis and, ultimately, to identify credible molecular targets for therapeutic intervention. As outlined in chapter 1.4.4, GWAS have now identified many variants associated with risk for psychiatric disorders. However, there is now a significant bottleneck for researchers in terms of translating these genetic findings into clinical outcomes. For example, as most risk loci reside in poorly characterised non-coding regions, it is a challenge for researchers to ascertain exactly how risk is conferred by these loci. Furthermore, with linkage disequilibrium making it difficult to isolate causal from associated variants, and with the spread, and volume, of associated loci across the genome, the financial and practical reserves required to examine every genetic lead individually would be astronomical. For these reasons, prioritisation of the genetic risk loci that are most likely to confer risk is essential.

From an initial list of 390 LOAD common variants, representing all the genome-wide significant index SNPs from the IGAP GWAS (80) and associated SNPs with an $R^2 \geq 0.8$, 29 SNPs were found to overlap microglial open chromatin regions. Both rs9381562 and rs28834970 were prioritised over the other 29 LOAD common variants for several reasons: (A) they had the largest allele-specific effects on motif binding potential, (B) evidence in the literature suggested that they were located in active regulatory regions in microglia, and (C) the motifs that were predicted to be disrupted at both loci had phenotypic relevance in microglia. As (A) and (B) were

covered in earlier sections, they will be discussed only briefly here; the majority of the following discussion will focus on (C).

The MotifbreakR analysis (Table 5.3) predicted that alleles at rs9381562 and rs28834970 were most disruptive to the DNA binding potential. Whilst this finding alone may be considered justification enough for prioritising these two SNPs, the fact that alleles at these loci disrupted motifs known to bind microglial lineage-determining transcription factors, and that these factors have also been linked with LOAD risk, suggested that these SNPs were excellent candidates for the EMSA.

For example, protein C-ets-1 (ETS1), is a member of the ETS family of transcription factors that bind to the 5' – GGA(A/T) – 3' motif, which was predicted to be disrupted by rs9381562. There are 28 transcription factors known to bind to the ETS domain which can be subdivided into 4 subgroups based on common, preferential interactions between the TF amino acids and nucleotides at position 4 of the motif (A/T), or at nucleotides directly flanking the motif (350). At the molecular level, ETS factors can inhibit or enhance DNA binding of cofactors to regulate transcription. In the brain, ETS binding factors have been implicated in neuronal pathfinding and differentiation, but these factors influence a wide range of biological processes across several tissue types including the skin and reproductive organs (351,352). One family that bind to the ETS domain are the SPI subfamily of proteins that includes the myeloid lineage determining transcription factor PU.1. Moreover, whilst over half the ETS family of transcription factors are ubiquitously expressed over several cell types, high expression of SPI1, which is the gene that encodes PU.1, is specific to myeloid cells (352). With regard to LOAD, a recent genome-wide study reported that a protective genome-wide significant LOAD risk SNP located in the 3' untranslated region of the SPI1 gene lowered the expression of SP1 in myeloid cells, and suggested that reduction of SPI1 expression levels could lead to partial amelioration of AD risk through modulation of myeloid cell gene expression (250).

The CCAAT/enhancer binding protein (C/EPB) family of transcription factors have 6 isoforms that bind to the consensus sequence 5'-CCAAT-3', predicted to be disrupted by rs28834970. The best characterised isoforms are C/EBP-α and C/EBP-β, which are ubiquitously expressed throughout the body and are primarily involved with cellular proliferation and differentiation (353). In the brain, CEBP-β has been implicated in neuroinflammatory regulation, where it has been shown to be

upregulated in activated microglia and astrocytes (354). Moreover, C/EBP-β is significantly increased in the cortex of LOAD patients compared with non-demented controls (355), and over-expression of C/EBP-β in LOAD mouse models accelerates LOAD-like symptoms, such as cognitive impairment, via modulation of delta-secretase transcription (356).

The case for selecting rs9381562 and rs28834970 was further strengthened by data from publicly available sources in section 5.3.2. (see figures 5.2 and 5.3) which demonstrated that loci encompassing either rs9381562 or rs28834970 associate with chromatin annotations that correlate with the presence of active enhancers in microglia. Clarifying that non-coding variants lie in genomic regions that are functionally relevant in a particular cell type is important as the chromatin landscape of every cell type is unique. So, for example, whilst an individual's genotype may show that they carry the LOAD risk allele at rs9381562, if rs9381562 is located in a microglial heterochromatic region, a region where chromatin is closed, any risk conferred by this SNP could not be mediated via this cell type. In effect, the region containing the SNP would, in terms of gene regulation, be silent. Regardless of whether the relevant transcription factors were expressed in microglia or not, the tightly packed histones at this site would sterically impede transcription factors, preventing them from accessing the underlying motifs that they bind to. Presented with the evidence described above, rs9381562 and rs28834970 were deemed suitable candidates for the EMSA assay.

By showing in section 5.3.3. figure 5.4, that oligonucleotides designed to include rs9381562 and rs28834970 alleles were capable of forming DNA-protein complexes with factors from BV2 nuclear extracts, I have established that the regions encompassing both SNPs are likely to be regulatory elements. Whilst the EMSA cannot directly indicate the nature of these putative regulatory elements (i.e. by identifying the elements' nature and/or status), these results do concord with evidence from the literature (see section 5.3.2) that these SNPs fall within active, enhancer regions. Moreover, as the DNA-protein interactions at each polymorphism can be distinguished in an allele-specific manner (see figure 5.4), it is inferred that allelic variants at these loci have functional relevance to the binding of transcription factors expressed by microglia. With regard to LOAD, this allele-specific effect could mediate all, or part, of the genetic association reported in the LOAD GWAS studies

for these loci, but further, more detailed, experiments will be required to determine this (see section 5.4.1).

In the PU.1 supershift assay (Figure 5.6), there was no evidence of a supershifted banding pattern; however, the allele specific change in band intensity between lanes containing the PU.1 antibody and IgG antibody may imply that PU.1 is present, if not directly detectable via EMSA, within this complex. It is plausible that if PU.1 is the critical DNA-binding factor that underpins the interaction of the entire protein complex, then the inclusion of the antibody may interfere with biotinylated DNA molecules forming DNA-protein interactions. This could explain why there is reduction in band intensity in well 7 compared with well 2, rather than a supershifted banding pattern. However, further experiments will be required to confirm this.

Although position weight matrices (PWMs) are used extensively to explore consensus sequence variation in humans, they do have limitations. First, PWMs assume that the binding energy of each nucleotide within a given motif is independent. Whilst, generally, this does appear to be a good model, as it models the hydrogen bonding that occurs between individual nucleotides and the amino acid side chains of transcription factors (357), it has been shown to be an over-simplification. For example, the orientation of the minor groove of the DNA backbone also influences transcription factor binding. Conformational changes in DNA structure can alter the electrostatic environment surrounding a protein complex and the DNA it binds to. This, in turn, can impact of the binding efficiency of the individual factors of the complex (358). As such, a point change in one motif can alter the electrostatic potential of the base, and surrounding environment, potentially altering DNA conformation. This could have a non-independent effect on the binding efficiency of factors elsewhere in the network (115). Interestingly, it has been suggested that DNA conformation is a genomic chromatin feature under evolutionary constraint in certain genomic regions, and that these regions correlated strongly with enhancer elements (359). Moreover, PWMs do not capture the binding characteristics of multiple identical TFs binding at a motif as a regulatory unit (i.e. dimers or trimers) as they only consider each transcription factors binding efficiency in isolation. Understanding the complexity of the gene regulatory machinery is a major challenge for the field of functional genomics, as is accurately modelling a poorly understood system.

Although the EMSA is inexpensive, relatively quick to perform, and provides a robust method to identify allele-specific effects impacting DNA-protein binding, it is an artificial system. The EMSA cannot identify the individual proteins forming the transcription factor complexes that bind to DNA, or provide information on exactly where along the oligonucleotide these complexes are bound (360,361). Moreover, as the DNA input for the assay is a synthetic, relatively short oligonucleotide, the chromatin environment that the DNA-protein interactions would take place within, *in vivo*, has not been properly modelled. As mentioned in chapter 1.3.2., *in vivo*, DNA is intertwined with histone proteins to form chromatin, and the conformation and post-translational modification of chromatin can influence context dependent transcription factor binding. As no histones have been added in the assay, any potential effects that chromatin conformation (or modification) would have on protein interactions have not been measured. That said, the EMSAs utility is in its simplicity as it provides a means to identify functional SNPs relatively quickly without the need for complicated, and expensive, genetic or cellular manipulation.

5.4.1 Future Work

Given the inconclusive result of the supershift assay, and the fact that transcription factors rarely bind to regulatory elements in isolation, future experiments should aim to characterise the proteins that form the complexes at both loci. One method that could be used for this is the DNA-affinity precipitation assay (DAPA). In this assay, oligonucleotides containing LOAD risk alleles are incubated with nuclear lysates to capture both DNA bound factors, and factors that associate with them, and the DNA-protein complexes are precipitated from the lysate using microbeads (362). These complexes are then profiled by mass spectrometry to identify the proteins present and compared to assess if any allele-specific differences can be detected.

Considering that there were clear allele-specific differences in DNA-protein binding at both loci, assessing the impact, if any, that variation at these sites has on gene expression is important. The gene editing technology CRISPR/Cas9 could be used to genetically modify microglial cell lines such that separate stable lines are generated for both allelic variants at each locus. CRISPR/Cas9 uses an RNA guidance system to target specific regions of DNA and then recruit DNA editing enzymes to modify the nucleotide sequence (363,364). Once stable lines are created, it would be possible to use a number of techniques to assess the degree to which differential binding of

nuclear factors impacts gene expression. For example, chromatin capture technologies, such as 3C and Hi-C, could be employed to assess any differences in promoter-enhancer interactions caused by allelic variation. This would also be a means to identify the target genes of each regulatory element (365). Furthermore, RNA-seq could be used to assess any effect allelic variation has on the expression of the putative regulated gene as well as on downstream gene expression.

Whilst a bioinformatics approach was used here for SNP prioritisation other methods exist such as expression quantitative trait loci (eQTL) analysis, which leverages genotypic and transcriptomic data to test for association of genetic loci with alterations in the levels of tissue specific gene expression (a locus associated with an alteration of gene expression is termed an eQTL). As such, this method provides the additional benefit of mapping non-coding genetic variants to the gene/s they putatively influence. The Genotype-Tissue Expression (GTEx) project is large-scale eQTL initiative which, to date, has assessed the association of genetic loci with gene expression in 54 adult human tissues (366), including several brain regions, and provides an open access repository to query eQTL in these tissues (https://www.gtexportal.org/home/). Interestingly, when querying these datasets, rs28834970 was reported to modify PTK2β gene expression in five tissues with the top hit in whole blood ($P = 1.2 \times 10^{-16}$), but not in brain, whilst rs9381562 was an eQTL in 26 human tissues, including 6 brain regions, the majority of which correlated with an alteration in expression of RP11-385F7.1. As these analyses were carried out in bulk tissue, it is not possible to assess how individual cell types contribute to the eQTL signal. Proportionally, microglia constitute 0.5-16% of the total glial population in the brain depending on the brain region considered (136), so it is likely that microglial specific eQTL signals are modified in bulk brain studies by competing or complimentary signals deriving from other brain cell-types. As such, there is a need for single cell eQTL studies which focus on individual cell types of the brain. At the time of writing, no microglial-specific eQTL analyses exist in the literature.

5.4.2 Concluding remarks

Using EMSA, I have investigated the functional consequences of two LOAD associated polymorphisms and shown that DNA-protein interactions associated with rs9381562 and rs22834970 are impacted in an allele-specific manner. This adds to the growing body of evidence that LOAD associated variants have functional

relevance in microglial cell lines, and that microglial dysfunction contributes to LOAD risk.

6 General Discussion

The primary aim of my thesis was to investigate whether gene regulatory processes in human microglia contribute to genetic risk for complex brain disorders. With this in mind, I was seeking to answer several questions (A) Could common variant associations with complex brain disorders be attributed to gene regulatory mechanisms in a specific cell type of the brain (i.e. microglia)? (B) If so, does the developmental time point at which primary microglia were isolated have an impact on this risk attribution? (C) If cell-specific gene regulatory mechanisms are impacted by brain disorder common risk variants, is it possible to discriminate causal, from associated, SNPs and identify genuine molecular perturbations that are caused by allelic variation? (D) Could microglial human-derived cell models recapitulate the open chromatin landscape of primary microglia?

In chapter 2, using stratified linkage disequilibrium regression analysis, I tested whether human adult *ex vivo* microglia open chromatin regions (OCRs) were enriched for SNP heritability associated with 7 complex brain disorders. I then repeated this test using adult neuronal OCRs derived from the ventrolateral pre-frontal cortex to test whether brain disorder SNP heritability was generic across two brain cell types or could be attributed more specifically to microglia. Of the 7 disorders tested, adult microglia OCRs were enriched for LOAD SNP heritability, capturing >50% of the total common variant component of the disorder, whilst adult neuronal OCRs were not significantly enriched. These data imply that LOAD-associated common variants operate in adult microglial OCRs and that a proportion the LOAD risk signal is mediated by immune cells. Due to the extensive overlap of the chromatin, and gene expression, profiles of microglia and peripheral myeloid cells (251) it was not possible to definitively rule out the involvement of peripheral immune cells, or to rule out involvement of other brain cell types, in LOAD common variant risk. Further evidence implicating microglial function in LOAD genetic risk was provided in chapter 4, when, using GARFIELD, I tested adult *ex vivo* microglial OCRs for enrichment of SNPs at two GWAS significance thresholds. Similar to the sLDSC analysis, LOAD-associated SNPs that were significant at the $P < 1 \times 10^{-5}$ threshold were enriched in adult MG OCRs, suggesting that the LOAD SNP enrichment signal is robust as it is replicable using independent statistical methodologies. Overall, these data indicate that substantial proportion of common variant risk for LOAD operates via gene regulatory

processes in adult microglia. These data are therefore consistent with other recent epigenomic studies supporting a primary role for microglia in this disease (367,368).

Having demonstrated that microglial-specific regulatory regions were implicated in LOAD common variant risk I next explored whether it was possible to partition LOAD SNP heritability further and test if it colocalised within OCRs containing specific transcription factor (TF) motifs. In chapter 2.3.4, by showing that LOAD SNP heritability was localised in OCRs containing motifs of key regulators of cell identity (*SPI1, RUNX1 and C/EBP*; 143,145,282,362), this strongly implies that immune cell-specific regulatory networks mediate LOAD common variant risk.

Pioneer factors such as PU.1 (encoded by *SPI1*) and CEBP-α, have been shown to act in collaboration to establish a core suite of myeloid cell enhancers, and their associated histone modifications, in macrophages (255,370). It is hypothesised that macrophages differentially activate specific subsets of these core enhancers in a tissue- and context- specific manner and that this is driven by environmental signals. These signals activate signal-dependant transcription factors which, in turn, interact with pioneer factors at selected enhancers to modulate gene expression. In microglia, TGFβ signalling and the transcription factor SMAD3 contribute to microglial-specific enhancer selection (139). Establishing whether there are brain-specific microglia enhancer patterns will be important to assess whether microglia-specific enhancers are uniquely enriched for LOAD risk SNPs. As such, it will be necessary to integrate open chromatin data with ChIP-seq data to determine whether core enhancers in microglia are inactive, active or poised (368). However, given that microglial chromatin features are likely to be altered by context-specific stimuli, functional data obtained from microglia in specific activation states and/or from specific brain regions is required to explore this fully. With regard to the cooperative binding of PU.1 or CEBP-α, it has been reported that mutations within the motifs of either of these factors result in the loss of binding of <u>both</u> transcription factors indicating that pioneer factor collaboration is critical for enhancer selection (370). Therefore, LOAD risk variants altering the DNA template in microglia OCRs containing these key factors, as reported here, have the potential to directly, and/or indirectly, disrupt the binding of pioneer factors and signal dependent transcription factors (370). As, *SPI1* has been identified as a LOAD risk gene these data provide further evidence of this gene's involvement in the disorder (208).

The purpose in annotating the genome for integration with GWAS data is to identify and prioritise the cell-types, regulatory processes and risk variants that drive the pathophysiology of brain disorders. Having shown that LOAD risk variants plausibly operate in microglial regulatory regions I next attempted to identify putatively causal LOAD risk SNPs, that overlapped microglial OCRs, to assess how these variants impacted transcription factor binding at the molecular level. As described in chapter 5, I prioritised 2 LOAD risk SNPs (rs9381562 and rs28834970) that overlapped adult *ex vivo* microglial open chromatin sites for molecular analysis based on allele-specific disruption of transcription factor binding potential at these loci. Using electrophoresis mobility shift assays (EMSAs), I demonstrated that DNA-protein binding was disrupted in an allele-specific manner at both loci, and that allele-specific effects at these loci resulted in >50% differences in DNA-protein binding (see table 5.4). These data demonstrate the benefit in functionally annotating GWAS data before experimental interrogation. When analysed separately, GWAS and chromatin accessibility data provide a holistic view of individual aspects of the genome (i.e. genomic loci that are implicated in brain disorder risk or regions that are involved in cell-specific gene regulation, respectively). In isolation these data are informative; however, when these datasets were combined, it was possible to focus global measures of LOAD risk to implicate specific cell types and specific molecular processes in risk burden.

Having measured enrichment of common risk variants in adult microglial regulatory regions, I next explored whether foetal microglial regulatory sites were similarly enriched (see chapter 4.3.3). Of the 8 traits tested, bipolar disorder (BPD) and schizophrenia (SCZ) associated SNPs (at the $P < 1 \times 10^{-5}$ GWAS threshold) were enriched in foetal microglial OCRs. Interestingly, it has been widely reported that there is significant genetic (and neurobiological) comorbidity between BPD and SCZ (371–375), so although additional work would be required to assess whether similar regulatory process were perturbed by disease-specific risk SNPs in foetal microglia, these data suggest that mechanisms underlying genetic risk for both disorders may operate in microglia during pre-natal neurodevelopment. It has long been theorised that schizophrenia has a neurodevelopmental component (2–5,376). However, as the GARFIELD analysis of brain disorder GWAS SNP enrichment in adult ex *vivo* microglia demonstrated (see chapter 4.3.5) that SCZ and BPD risk SNPs were also enriched in adult microglia, this implies that these variants have the potential to impact microglial function throughout life. An intriguing hypotheses that could explain these findings with regard to schizophrenia, is the ‘2-hit’ model which proposes that

schizophrenia symptoms are caused by two brain insults occurring during discrete critical periods of vulnerability (before birth and during adolescence when extensive neuronal rewiring is occurring; 4,368,369). In this model, abnormal neuronal growth is hypothesised to occur in specific neuronal networks before birth that elicit the premorbid cognitive deficits seen in pre-psychotic individuals. Then, during the brain maturation process in adolescence, excessive synaptic elimination occurs and this interacts cumulatively with the mechanisms perturbed earlier in life to induce psychosis. As microglia are involved in both in promoting neuronal growth and synapse elimination (379,380), it is possible that genetic perturbations to microglial gene regulatory processes during key periods of vulnerability could increase liability to this condition.

Given the need for reliable human cell models of microglia for brain disorder research, it is important to confirm how well *in vitro* cell lines recapitulate the regulatory profile of primary, or *in vivo,* cells. In chapter 3, I reported that iPSC-derived microglia (iPSC_MG) and iPSC-derived macrophage precursor (IPSC_MΩpre) OCRs were enriched for motifs recognised by the myeloid-specific transcription factors Spi1, C/EBP-α and IRF8 (144,146,285,369). As these were similar to the motifs found to be enriched in adult *ex vivo* microglial OCRs (see section 2.3.1), these results indicated that both iPSC lines had a myeloid-like transcription factor motif profile that was similar to adult *ex vivo* microglia. However, it was not possible to distinguish between the iPSC lines using this measure. In the principal component analysis (PCA; see figure 3.2) the distinctive separation of myeloid and lymphoid cells at 0.0 on the y-axis, suggested that the open chromatin landscapes of all *in vitro* cell lines were myeloid-like. However, as the iPSC line samples were clustered more closely with samples from blood monocytes than adult *ex vivo* microglia on PC2, this implied that the addition of IL-34 during the final stage of iPSC differentiation protocol was not sufficient to skew the chromatin landscape of iPSC_MΩpres toward that of adult *ex vivo* microglia. Interestingly, when adding the foetal microglial ATAC-seq data to the PCA (see figure 4.3), samples from both iPSC lines clustered more closely with the foetal microglial samples on PC2 than samples from either adult *ex vivo* microglia or the peripheral blood monocytes. Relating chromatin features in iPSC-derived microglia to that of primary foetal cells may be a better means of assessing how well the chromatin landscape of the iPSC lines recapitulate the *in vivo* state than using primary adult cells. Indeed, all groups that have produced iPSC-derived microglial protocols recently use gene expression in human foetal microglia as their gold standard (268,277–279,295). However, regardless of the primary cell type used,

these data imply that there is no clear global distinction in the chromatin profile between the iPSC_MΩpres and the iPSC_MG, at least sufficient enough to claim that the IPSC_MG are more ‘microglial-like' than the iPSC_MΩpre.

The generation of iPSC-derived microglia is still a relatively new methodology and there is, as yet, no standardised protocol for this procedure. For example, in 2017, 4 groups released bespoke protocols that used differing cell seeds (embryonic stem cells and or iPSCs), surface markers for cell isolation, differentiation factors and incubation periods to generate microglia-like cells. As such, it is necessary to thoroughly validate these methods and determine which, if any, most reliably models the *in vivo* state (as it is unlikely that multiple methods will be sustainable due to potential confounds that arise when comparing results using alternate approaches). Although the use of iPSC-derived microglia for brain disorder research does provide enticing benefits when compared to using primary microglia (i.e. they are accessible, abundant, easy to produce and can be patient-derived), several key issues have still to be addressed before they are widely accepted as an appropriate model. For example, it is not yet known how to maintain an *in vivo* like gene expression signature in cultured cells (265,266,280,381). Some progress has been made which has informed iPSC protocols recently; for example, it has been reported that the application of CSF1 and TGFβ partially rescues the *in vivo* gene expression signature of microglia *in vitro* (382), but full recapitulation has yet to be realised. Moreover, as iPSC-derived microglia are created *in vitro,* and have never been exposed to the brain microenvironment, it is unclear whether they can ever fully recapitulate the phenotype/s of microglia *in vivo* (270). Questions remain about whether the ontological derivation of microglia intrinsically impacts the range of functions that fully differentiated cells are capable of (266,383) and whether iPSC protocols that mimic mouse microglial differentiation apply to human microglial development (268). On a positive note, it has been shown that when *ex vivo* microglia placed (and manipulated) in culture are reintroduced to the brain they swiftly regain their homeostatic *in vivo* gene expression profile (236,266). This has promising implications for microglial cell-based therapy as it suggests that *ex vivo* (and potentially patient-derived iPSC) microglia can be cultured, modified and transplanted, and that the brain is already equipped to rescue the microglial phenotype. However, the success of such therapies will be predicated on identifying the differentiation factors that are required to fully induce homeostatic microglial phenotypes in culture. If a reliable iPSC-derived microglial model can be obtained, it’s use in combination with functional data such as that presented here (i.e. the identification of a functional GWAS risk SNP in a

microglial specific OCR) paves the way for accurate modelling of molecular risk mechanisms *in vitro* using gene editing techniques such as CRISPR (363,364).

It is an exciting time for brain disorder research as advances in single cell analyses make it possible to discriminate between individual cell types of the brain, including heterogeneous cellular subtypes that may have distinct transcriptomic and epigenomic profiles in discrete brain regions (384,385). Indeed, microglia heterogeneity, and any relevance this has to health and disease, is a current interest in the field. Brain region dependent differences have been reported in microglia morphology (386), density (386) gene expression (387) and cell surface marker density (388) that are perhaps indicative of non-uniform brain microenvironments in which each distinct cell population resides (387). Measures of microglia longevity and turnover have also been reported as brain region dependent (389). In a recent microglial single cell analysis comparing the RNA-seq profile of cells extracted from 3 brain regions (dorsolateral pre-frontal cortex, hippocampus and temporal cortex) in 15 individuals, 14 microglial clusters were identified which showed differential expression of homeostatic, proliferative and interferon response genes (390). Microglia have also been reported as being sexually dysmorphic with morphological, functional and brain region dependant microglial sex differences being reported in mice (391). This has led to speculation that sex-specific microglial dysfunction may explain the sexual dimorphism seen in brain disorders such as LOAD and ASD (202,392). The identification of niche populations of microglia leads to important questions regarding their role in the aetiology of complex brain disorders. For example, in Alzheimer's disease, where pathology develops in the hippocampus before spreading more widely in the brain, do specific populations of hippocampal microglia have unique vulnerabilities which influence disease onset and progression (393)? The role that specific populations of microglia have in the brain, and how this relates to complex brain disorders, is still an open question.

Another interesting aspect of microglial biology which may impact the aetiologies of complex brain disorders is the role of systemic immune processes that influence microglial phenotype. For example, acute and chronic systemic infections, as indexed by a serum increase in the proinflammatory cytokine tumour necrosis factor α, have been reported to cause a 2-10 fold rate of increase of cognitive decline in patients with LOAD respectively (394). It is proposed that this process is driven by a phenomenon called microglial 'priming' where, after an initial immune insult, long-term epigenetic changes occur in microglia which make them more sensitive, and

react more aggressively, to a second immune insult (395,396). Primed microglia upregulate genes associated with increased phagocytosis, antigen presentation and lysosomal activity (397) and several processes have been reported to induce microglial priming in humans such as aging, stress, amyloid-β protein aggregates and degenerating neurons (the latter due to the loss of neuronal ligands that inhibit microglial activation; 384). In LOAD it is hypothesised that either progressive increase of Amyloid-β, degenerating neurons or local ischaemia leads to microglia becoming permanently primed such that they swiftly become fully activated during a systemic infection thereby contributing to the progression of LOAD through increased production of reactive oxygen species (398,399). Indeed as chronic systemic disorders with a proinflammatory phenotype such as heart disease (400), obesity (401) and diabetes (402) are risk factors for LOAD, some investigators propose that culminative insults affecting primed microglia over time may be a causative factor in LOAD (394,403).

While epigenomic methods such as ATAC-Seq can be exploited to better understand the cellular basis of complex traits, it does not necessarily indicate which genes are functionally impacted by risk variation and how they are altered. In recent years, chromosome conformation capture (3C) -based methods have been developed in order to map distal interactions between enhancers and their regulated genes (404,405). Moreover, with sufficient sample sizes, it is possible to perform expression quantitative trait loci (eQTL) mapping studies, which identify genetic effects on gene expression on a genome-wide scale. While current eQTL studies are typically carried out in bulk tissue, it will be necessary for future investigations to focus on discrete cell populations, including microglia.

7 References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (5th Edition). American Journal of Psychiatry. 2013.
2. Weinberger DR. Implications of Normal Brain Development for the Pathogenesis of Schizophrenia. Arch Gen Psychiatry. 1987;
3. Gross G, Huber G. Is schizophrenia a neurodevelopmental disorder? Neurology Psychiatry and Brain Research. 1997.
4. Fatemi SH, Folsom TD. The neurodevelopmental hypothesis of Schizophrenia, revisited. Schizophrenia Bulletin. 2009.
5. Birnbaum R, Weinberger DR. Genetic insights into the neurodevelopmental origins of schizophrenia. Nature Reviews Neuroscience. 2017.
6. Rutter M. Research review: Child psychiatric diagnosis and classification: Concepts, findings, challenges and potential. Journal of Child Psychology and Psychiatry and Allied Disciplines. 2011.
7. Lord C, Elsabbagh M, Baird G, Veenstra-Vanderweele J. Autism spectrum disorder. Lancet [Internet]. 2018 Aug 11;392(10146):508–20. Available from: https://doi.org/10.1016/S0140-6736(18)31129-2
8. Lyall K, Croen L, Daniels J, Fallin MD, Ladd-Acosta C, Lee BK, et al. The Changing Epidemiology of Autism Spectrum Disorders. Annu Rev Public Health. 2017;
9. Elsabbagh M, Divan G, Koh YJ, Kim YS, Kauchali S, Marcín C, et al. Global Prevalence of Autism and Other Pervasive Developmental Disorders. Autism Res. 2012;
10. Luciano K. Autism spectrum disorder. J Am Acad PAs [Internet]. 2016;29(10). Available from: https://journals.lww.com/jaapa/Fulltext/2016/10000/Autism_spectrum_disorder.2.aspx
11. Polanczyk G, De Lima MS, Horta BL, Biederman J, Rohde LA. The worldwide prevalence of ADHD: A systematic review and metaregression analysis. Am J Psychiatry. 2007;
12. Faraone S V, Asherson P, Banaschewski T, Biederman J, Buitelaar JK, Ramos-Quiroga JA, et al. Attention-deficit/hyperactivity disorder.

Nat Rev Dis Prim [Internet]. 2015 Aug 6;1:15020. Available from: https://doi.org/10.1038/nrdp.2015.20

13. Simon V, Czobor P, Bálint S, Mészáros Á, Bitter I. Prevalence and correlates of adult attention-deficit hyperactivity disorder: meta-analysis. Br J Psychiatry [Internet]. 2018/01/02. 2009;194(3):204–11. Available from: https://www.cambridge.org/core/article/prevalence-and-correlates-of-adult-attentiondeficit-hyperactivity-disorder-metaanalysis/FBBDADEA596D69D26F49318ECAD410C4
14. Vieta E, Berk M, Schulze TG, Carvalho AF, Suppes T, Calabrese JR, et al. Bipolar disorders. Nat Rev Dis Prim [Internet]. 2018 Mar 8;4:18008. Available from: https://doi.org/10.1038/nrdp.2018.8
15. Cullen B, Ward J, Graham NA, Deary IJ, Pell JP, Smith DJ, et al. Prevalence and correlates of cognitive impairment in euthymic adults with bipolar disorder: A systematic review. J Affect Disord [Internet]. 2016;205:165–81. Available from: http://www.sciencedirect.com/science/article/pii/S0165032716307534
16. Jamison KR, Ph D. Manic-Depressive Illness Bipolar Disorders and Recurrent Depression. Manic-Depressive Illn Bipolar Disord Recurr Depress. 2007;
17. Merikangas KR, Jin R, He JP, Kessler RC, Lee S, Sampson NA, et al. Prevalence and correlates of bipolar spectrum disorder in the World Mental Health Survey Initiative. Arch Gen Psychiatry. 2011;
18. Prince, M, Knapp, M, Guerchet, M, McCrone, P, Prina, M, Comas-Herrera, A, Wittenberg, R, Adelaja, B, Hu, B, King, D, Rehill, A and Salimkumar D. Dementia UK: Second edition. Alzheimer's Society. 2014.
19. Scheff SW, Price DA, Schmitt FA, Mufson EJ. Hippocampal synaptic loss in early Alzheimer's disease and mild cognitive impairment. Neurobiol Aging. 2006;
20. Terry RD, Masliah E, Salmon DP, Butters N, DeTeresa R, Hill R, et al. Physical basis of cognitive alterations in alzheimer's disease: Synapse loss is the major correlate of cognitive impairment. Ann Neurol. 1991;
21. Spires-Jones TL, Hyman BT. The Intersection of Amyloid Beta and Tau at Synapses in Alzheimer's Disease. Neuron. 2014.

22. Holtzman DM, Morris JC, Goate AM. Alzheimer's disease: The challenge of the second century. Science Translational Medicine. 2011.
23. Mrak RE. Microglia in Alzheimer brain: A neuropathological perspective. International Journal of Alzheimer's Disease. 2012.
24. Otte C, Gold SM, Penninx BW, Pariante CM, Etkin A, Fava M, et al. Major depressive disorder. Nat Rev Dis Prim [Internet]. 2016 Sep 15;2:16065. Available from: https://doi.org/10.1038/nrdp.2016.65
25. Seedat S, Scott KM, Angermeyer MC, Berglund P, Bromet EJ, Brugha TS, et al. Cross-national associations between gender and mental disorders in the World Health Organization World Mental Health Surveys. Arch Gen Psychiatry. 2009;
26. Bromet E, Andrade LH, Hwang I, Sampson NA, Alonso J, de Girolamo G, et al. Cross-national epidemiology of DSM-IV major depressive episode. BMC Med. 2011;
27. Vos T, Barber RM, Bell B, Bertozzi-Villa A, Biryukov S, Bolliger I, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: A systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015;
28. Lahey BB. Public Health Significance of Neuroticism. Am Psychol. 2009;
29. Malouff JM, Thorsteinsson EB, Schutte NS. The relationship between the five-factor model of personality and symptoms of clinical disorders: A meta-analysis. J Psychopathol Behav Assess. 2005;
30. Khan AA, Jacobson KC, Gardner CO, Prescott CA, Kendler KS. Personality and comorbidity of common psychiatric disorders. Br J Psychiatry [Internet]. 2018/01/02. 2005;186(3):190–6. Available from: https://www.cambridge.org/core/article/personality-and-comorbidity-of-common-psychiatric-disorders/74AD511EDEDED8C0951F0054A4AD6D05
31. Perälä J, Suvisaari J, Saarni SI, Kuoppasalmi K, Isometsä E, Pirkola S, et al. Lifetime prevalence of psychotic and bipolar I disorders in a general population. Arch Gen Psychiatry. 2007;

32. Kahn RS, Sommer IE, Murray RM, Meyer-Lindenberg A, Weinberger DR, Cannon TD, et al. Schizophrenia. Nat Rev Dis Prim [Internet]. 2015 Nov 12;1:15067. Available from: https://doi.org/10.1038/nrdp.2015.67
33. Tripathi A, Kar SK, Shukla R. Cognitive deficits in schizophrenia: Understanding the biological correlates and remediation strategies. Clinical Psychopharmacology and Neuroscience. 2018.
34. Heinrichs RW, Zakzanis KK. Neurocognitive deficit in schizophrenia: A quantitative review of the evidence. Neuropsychology. 1998;
35. Häfner H, Riecher-Rössler A, Hambrecht M, Maurer K, Meissner S, Schmidtke A, et al. IRAOS: an instrument for the assessment of onset and early course of schizophrenia. Schizophr Res. 1992;
36. Kahn RS, Keefe RSE. Schizophrenia is a cognitive illness: Time for a change in focus. JAMA Psychiatry. 2013.
37. Laursen TM, Munk-Olsen T, Vestergaard M. Life expectancy and cardiovascular mortality in persons with schizophrenia. Current Opinion in Psychiatry. 2012.
38. McGrath J, Saha S, Chant D, Welham J. Schizophrenia: A concise overview of incidence, prevalence, and mortality. Epidemiologic Reviews. 2008.
39. Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a Complex Trait: Evidence from a Meta-analysis of Twin Studies. Arch Gen Psychiatry. 2003;
40. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, et al. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry [Internet]. 2006;63(2):168. Available from: http://archpsyc.jamanetwork.com/article.aspx?doi=10.1001/archpsyc.63.2.168
41. Henriksen MG, Nordgaard J, Jansson LB. Genetics of Schizophrenia: Overview of Methods, Findings and Limitations. Front Hum Neurosci. 2017;
42. Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: The emerging picture and its implications. Nature

Reviews Genetics. 2012.
43. Weinberger DR. Thinking about schizophrenia in an era of genomic medicine. American Journal of Psychiatry. 2019.
44. Bray NJ, O'Donovan MC. The genetics of neuropsychiatric disorders. Brain Neurosci Adv. 2018;
45. Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, et al. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature. 1991;
46. Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, Ikeda M, et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature. 1995;
47. Dai M-H, Zheng H, Zeng L-D, Zhang Y. The genes associated with early-onset Alzheimer's disease. Oncotarget. 2017;
48. Altmüller J, Palmer LJ, Fischer G, Scherb H, Wjst M. Genomewide Scans of Complex Human Diseases: True Linkage Is Hard to Find. Am J Hum Genet. 2002;
49. Risch NJ. Searching for genetic determinants in the new millennium. Nature. 2000.
50. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genetics in Medicine. 2002.
51. Bertram L, Lill CM, Tanzi RE. The genetics of alzheimer disease: Back to the future. Neuron. 2010.
52. Raychaudhuri S. Mapping rare and common causal alleles for complex human diseases. Cell. 2011.
53. Malhotra D, Sebat J. CNVs: Harbingers of a rare variant revolution in psychiatric genetics. Cell. 2012.
54. Rees E, Walters JTR, Georgieva L, Isles AR, Chambert KD, Richards AL, et al. Analysis of copy number variations at 15 schizophrenia-associated loci. Br J Psychiatry. 2014;
55. Kirov G. CNVs in neuropsychiatric disorders. Human Molecular Genetics. 2015.
56. Broomer A, Veltman JA, Simonic I, Schwartz CE, Bongers EMHF, Eichler EE, et al. Recurrent Rearrangements of Chromosome 1q21.1

and Variable Pediatric Phenotypes. N Engl J Med. 2008;
57. Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, et al. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat Genet. 2008;
58. Chen X, Shen Y, Zhang F, Chiang C, Pillalamarri V, Blumenthal I, et al. Molecular analysis of a deletion hotspot in the NRXN1 region reveals the involvement of short inverted repeats in deletion CNVs. Am J Hum Genet. 2013;
59. Williams NM, Zaharieva I, Martin A, Langley K, Mantripragada K, Fossdal R, et al. Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: A genome-wide analysis. Lancet. 2010;
60. Charney AW, Stahl EA, Green EK, Chen CY, Moran JL, Chambert K, et al. Contribution of Rare Copy Number Variants to Bipolar Disorder Risk Is Limited to Schizoaffective Cases. Biol Psychiatry. 2019;
61. Rucker JJH, Tansey KE, Rivera M, Pinto D, Cohen-Woods S, Uher R, et al. Phenotypic association analyses with copy number variation in recurrent depressive disorder. Biol Psychiatry. 2016;
62. Cuccaro D, De Marco EV, Cittadella R, Cavallaro S. Copy number variants in Alzheimer's disease. Journal of Alzheimer's Disease. 2016.
63. Veltman JA, Brunner HG. De novo mutations in human genetic disease. Nature Reviews Genetics. 2012.
64. Conrad DF, Keebler JEM, Depristo MA, Lindsay SJ, Zhang Y, Casals F, et al. Variation in genome-wide mutation rates within and between human families. In: Nature Genetics. 2011.
65. Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, Gormley P, et al. De novo mutations in schizophrenia implicate synaptic networks. Nature. 2014;506(7487):179–84.
66. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;
67. Goldstein ND, Tager-Flusberg H, Lee BK. Mapping collaboration networks in the world of autism research. Autism Res. 2015;
68. Sims R, Van Der Lee SJ, Naj AC, Bellenguez C, Badarinarayan N,

Jakobsdottir J, et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. Nat Genet. 2017;

69. Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, et al. TREM2 variants in Alzheimer's disease. N Engl J Med. 2013;368(2).
70. Jonsson T, Stefansson H, Steinberg S, Jonsdottir I, Jonsson P V, Snaedal J, et al. Variant of TREM2 associated with the risk of Alzheimer's disease. N Engl J Med. 2013;
71. Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'Ang LY, et al. The international HapMap project. Nature. 2003;
72. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;
73. Harold D, Abraham R, Hollingworth P, Sims R, Hamshere M, Pahwa JS, et al. Genome-Wide Association Study Identifies Variants at CLU and PICALM Associated with Alzheimer's Disease, and Shows Evidence for Additional Susceptibility Genes. Nat Genet. 2009;41(10):1088–93.
74. Lambert J-C, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet [Internet]. 2009;41(10):1094–9. Available from: http://dx.doi.org/10.1038/ng.439
75. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science (80-). 1993;
76. Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, Boada M, et al. Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA. 2010;303(18):1832–40.
77. Naj AC, Jun G, Beecham GW, Wamg L-S, Vardarajan BN, Buros J, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet. 2011;43:436–41.

78. Hollingworth P, Harold D, Sims R, Gerrish A, Lambert J-C, Carrasquillo MM, et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet [Internet]. 2011;43(5):429–35. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3084173&tool=pmcentrez&rendertype=abstract
79. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Naj AC, Boland A, et al. Meta-analysis of genetic association with diagnosed Alzheimer's disease identifies novel risk loci and implicates Abeta, Tau, immunity and lipid processing. bioRxiv. 2018;
80. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet [Internet]. 2013;45(12):1452–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3896259&tool=pmcentrez&rendertype=abstract
81. Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature [Internet]. 2014;511(7510):421–7. Available from: http://www.nature.com/doifinder/10.1038/nature13595
82. Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet. 2018;
83. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;
84. Stahl EA, Breen G, Forstner AJ, McQuillin A, Ripke S, Trubetskoy V, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. Nat Genet. 2019;
85. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018;

86. Ramanan VK, Shen L, Moore JH, Saykin AJ. Pathway analysis of genomic data: Concepts, methods, and prospects for future development. Trends in Genetics. 2012.
87. Jin L, Zuo XY, Su WY, Zhao XL, Yuan MQ, Han LZ, et al. Pathway-based analysis tools for complex diseases: A Review. Genomics, Proteomics and Bioinformatics. 2014.
88. Jones L, Lambert JC, Wang LS, Choi SH, Harold D, Vedernikov A, et al. Convergent genetic and expression data implicate immunity in Alzheimer's disease. Alzheimer's Dement. 2015;11(6).
89. Edwards SL, Beesley J, French JD, Dunning M. Beyond GWASs: Illuminating the dark road from association to function. American Journal of Human Genetics. 2013.
90. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science (80-). 2012;
91. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. American Journal of Human Genetics. 2018.
92. Nishizaki SS, Boyle AP. Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms. Trends in Genetics. 2017.
93. Feingold EA, Good PJ, Guyer MS, Kamholz S, Liefer L, Wetterstrand K, et al. The ENCODE (ENCyclopedia of DNA Elements) Project. Science. 2004.
94. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. 2012;
95. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. Nature Biotechnology. 2010.
96. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;
97. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature.

1997;
98. Robinson PJ, Rhodes D. Structure of the “30 nm” chromatin fibre: A key role for the linker histone. Current Opinion in Structural Biology. 2006.
99. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;
100. Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nature Reviews Genetics. 2016.
101. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nature Reviews Genetics. 2019;
102. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol. 2015;2015:21.29.1-21.29.9.
103. Kimura K, Wakamatsu A, Suzuki Y, Ota T, Nishikawa T, Yamashita R, et al. Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. Genome Res. 2006;
104. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome Res. 2006;
105. Maston GA, Evans SK, Green MR. Transcriptional Regulatory Elements in the Human Genome. Annu Rev Genomics Hum Genet. 2006;
106. Landry JR, Mager DL, Wilhelm BT. Complex controls: The role of alternative promoters in mammalian genomes. Trends in Genetics. 2003.
107. Xin D, Hu L, Kong X. Alternative promoters influence alternative splicing at the genomic level. PLoS One. 2008;
108. Kamat A, Hinshelwood MM, Murry BA, Mendelson CR. Mechanisms in tissue-specific regulation of estrogen biosynthesis in humans. Trends in Endocrinology and Metabolism. 2002.
109. Bonham K, Ritchie SA, Dehm SM, Snyder K, Boyd FM. Alternative, human SRC promoter and its regulation by hepatic nuclear factor-1α. J

Biol Chem. 2000;
110. Saleh A, Makrigiannis AP, Hodge DL, Anderson SK. Identification of a Novel Ly49 Promoter That Is Active in Bone Marrow and Fetal Thymus. J Immunol. 2014;
111. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;
112. Kadauke S, Blobel GA. Chromatin loops in gene regulation. Biochimica et Biophysica Acta - Gene Regulatory Mechanisms. 2009.
113. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;
114. Cowie P, Hay EA, Mackenzie A. The noncoding human genome and the future of personalised medicine. Expert Reviews in Molecular Medicine. 2015.
115. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. 2013;
116. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018;
117. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: Function, expression and evolution. Nature Reviews Genetics. 2009.
118. Consortium EP, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C a, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2013;489(7414):57–74.
119. Mayran A, Drouin J. Pioneer transcription factors shape the epigenetic landscape. Journal of Biological Chemistry. 2018.
120. Zhang DX, Glass CK. Towards an understanding of cell-specific functions of signal-dependent transcription factors. J Mol Endocrinol. 2013;
121. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;
122. Carotta S, Wu L, Nutt SL. Surprising new roles for PU.1 in the adaptive immune response. Immunol Rev. 2010;

123. Barozzi I, Simonatto M, Bonifacio S, Yang L, Rohs R, Ghisletti S, et al. Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers. Mol Cell. 2014;
124. Kouzarides T. Chromatin Modifications and Their Function. Cell. 2007.
125. Kouzarides T. Chromatin modifications and their function. Cell. 2007;
126. Tessarz P, Kouzarides T. Histone core modifications regulating nucleosome structure and dynamics. Nat Rev Mol Cell Biol. 2014;
127. Greer EL, Shi Y. Histone methylation: A dynamic mark in health, disease and inheritance. Nature Reviews Genetics. 2012.
128. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. Cell. 2007;
129. Bernstein BE, Humphrey EL, Erlich RL, Schneider R, Bouman P, Liu JS, et al. Methylation of histone H3 Lys 4 in coding regions of active genes. Proc Natl Acad Sci. 2002;
130. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007;
131. Park PJ. ChIP-seq: Advantages and challenges of a maturing technology. Nature Reviews Genetics. 2009.
132. Edwards JR, Yarychkivska O, Boulard M, Bestor TH. DNA methylation and DNA methyltransferases. Epigenetics and Chromatin. 2017.
133. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nature Reviews Genetics. 2010.
134. Liu B, Montgomery SB. Identifying causal variants and genes using functional genomics in specialized cell types and contexts. Hum Genet. 2019;
135. Fullard JF, Hauberg ME, Bendl J, Egervari G, Cirnaru MD, Reach SM, et al. An atlas of chromatin accessibility in the adult human brain. Genome Res. 2018;
136. Mittelbronn M, Dietz K, Schluesener HJ, Meyermann R. Local distribution of microglia in the normal adult human central nervous system differs by up to one order of magnitude. Acta Neuropathol.

2001;
137. P. DR-H. Cytology and Cellular Pathology of the Nervous System. Arch Intern Med. 1932;
138. Nayak D, Roth TL, McGavern DB. Microglia Development and Function. Annu Rev Immunol. 2014;
139. Gosselin D, Link VM, Romanoski CE, Fonseca GJ, Eichenfield DZ, Spann NJ, et al. Environment drives selection and function of enhancers controlling tissue-specific macrophage identities. Cell. 2014;
140. Lavin Y, Winter D, Blecher-Gonen R, David E, Keren-Shaul H, Merad M, et al. Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. Cell. 2014;
141. Matcovitch-Natan O, Winter DR, Giladi A, Aguilar SV, Spinrad A, Sarrazin S, et al. Microglia development follows a stepwise program to regulate brain homeostasis. Science (80-). 2016;
142. Hammond TR, Robinton D, Stevens B. Microglia and the Brain: Complementary Partners in Development and Disease. Annu Rev Cell Dev Biol. 2018;
143. Ginhoux F, Prinz M. Origin of microglia: Current concepts and past controversies. Cold Spring Harb Perspect Biol. 2015;
144. Ginhoux F, Greter M, Leboeuf M, Nandi S, See P, Gokhan S, et al. Fate mapping analysis reveals that adult microglia derive from primitive macrophages. Science (80-). 2010;
145. Schulz C, Perdiguero EG, Chorro L, Szabo-Rogers H, Cagnard N, Kierdorf K, et al. A Lineage of Myeloid Cells Independent of Myb and Hematopoietic Stem Cells. Science. 2012;
146. Kierdorf K, Erny D, Goldmann T, Sander V, Schulz C, Perdiguero EG, et al. Microglia emerge from erythromyeloid precursors via Pu.1-and Irf8-dependent pathways. Nat Neurosci. 2013;
147. Monier A, Evrard P, Gressens P, Verney C. Distribution and differentiation of microglia in the human encephalon during the first two trimesters of gestation. J Comp Neurol. 2006;
148. Monier A, Adle-Biassette H, Delezoide A-L, Evrard P, Gressens P, Verney C. Entry and Distribution of Microglial Cells in Human Embryonic and Fetal Cerebral Cortex. J Neuropathol Exp Neurol.

2007;
149. Andjelkovic AV, Nikolic B, Pachter JS, Zecevic N. Macrophages/microglial cells in human central nervous system during development: an immunohistochemical study. Brain Res. 1998;
150. Ajami B, Bennett JL, Krieger C, Tetzlaff W, Rossi FM V. Local self-renewal can sustain CNS microglia maintenance and function throughout adult life. Nat Neurosci. 2007;
151. Askew K, Li K, Olmos-Alonso A, Garcia-Moreno F, Liang Y, Richardson P, et al. Coupled Proliferation and Apoptosis Maintain the Rapid Turnover of Microglia in the Adult Brain. Cell Rep. 2017;
152. Ajami B, Bennett JL, Krieger C, McNagny KM, Rossi FM V. Infiltrating monocytes trigger EAE progression, but do not contribute to the resident microglia pool. Nat Neurosci. 2011;
153. Hashimoto D, Chow A, Noizat C, Teo P, Beasley MB, Leboeuf M, et al. Tissue-resident macrophages self-maintain locally throughout adult life with minimal contribution from circulating monocytes. Immunity. 2013;
154. Réu P, Khosravi A, Bernard S, Mold JE, Salehpour M, Alkass K, et al. The Lifespan and Turnover of Microglia in the Human Brain. Cell Rep. 2017;
155. Beers DR, Henkel JS, Xiao Q, Zhao W, Wang J, Yen AA, et al. Wild-type microglia extend survival in PU.1 knockout mice with familial amyotrophic lateral sclerosis. Proc Natl Acad Sci. 2006;
156. Cronk JC, Filiano AJ, Louveau A, Marin I, Marsh R, Ji E, et al. Peripherally derived macrophages can engraft the brain independent of irradiation and maintain an identity distinct from microglia. J Exp Med. 2018;
157. Bennett FC, Bennett ML, Yaqoob F, Mulinyawe SB, Grant GA, Hayden Gephart M, et al. A Combination of Ontogeny and CNS Environment Establishes Microglial Identity. Neuron. 2018;
158. Prinz M, Priller J, Sisodia SS, Ransohoff RM. Heterogeneity of CNS myeloid cells and their roles in neurodegeneration. Nature Neuroscience. 2011.
159. Prinz M, Priller J. Tickets to the brain: Role of CCR2 and CX3CR1 in myeloid cell entry in the CNS. J Neuroimmunol. 2010;

160. Hsiao EY, McBride SW, Chow J, Mazmanian SK, Patterson PH. Modeling an autism risk factor in mice leads to permanent immune dysregulation. Proc Natl Acad Sci. 2012;
161. Derecki NC, Cronk JC, Lu Z, Xu E, Abbott SBG, Guyenet PG, et al. Wild-type microglia arrest pathology in a mouse model of Rett syndrome. Nature. 2012;
162. Priller J, Prinz M. Targeting microglia in brain disorders. Science. 2019.
163. Nimmerjahn A, Kirchhoff F, Helmchen F. Neuroscience: Resting microglial cells are highly dynamic surveillants of brain parenchyma in vivo. Science (80-). 2005;
164. Tremblay MĚ, Lowery RL, Majewska AK. Microglial interactions with synapses are modulated by visual experience. PLoS Biol. 2010;
165. Wake H, Moorhouse AJ, Miyamoto A, Nabekura J. Microglia: Actively surveying and shaping neuronal circuit structure and function. Trends in Neurosciences. 2013.
166. Lehnardt S. Innate immunity and neuroinflammation in the CNS: The role of microglia in toll-like receptor-mediated neuronal injury. GLIA. 2010.
167. Chao CC, Hu S, Molitor TW, Shaskan EG, Peterson PK. Activated microglia mediate neuronal cell injury via a nitric oxide mechanism. J Immunol. 1992;
168. Boje KM, Arora PK. Microglial-produced nitric oxide and reactive nitrogen oxides mediate neuronal cell death. Brain Res. 1992;
169. Franco R, Fernández-Suárez D. Alternatively activated microglia and macrophages in the central nervous system. Progress in Neurobiology. 2015.
170. Cherry JD, Olschowka JA, O'Banion MK. Neuroinflammation and M2 microglia: The good, the bad, and the inflamed. Journal of Neuroinflammation. 2014.
171. Brockhaus J, Möller T, Kettenmann H. Phagocytozing ameboid microglial cells studied in a mouse corpus callosum slice preparation. Glia. 1996;
172. Petersen MA, Dailey ME. Diverse Microglial Motility Behaviors during Clearance of Dead Cells in Hippocampal Slices. Glia. 2004;

173. Martinez FO, Gordon S. The M1 and M2 paradigm of macrophage activation: time for reassessment. F1000Prime Rep. 2014;
174. Ransohoff RM. A polarizing question: Do M1 and M2 microglia exist. Vol. 19, Nature Neuroscience. 2016.
175. Morganti JM, Riparip LK, Rosi S. Call off the dog(ma): M1/M2 polarization is concurrent following traumatic brain injury. PLoS One. 2016;
176. Kim CC, Nakamura MC, Hsieh CL. Brain trauma elicits non-canonical macrophage activation states. J Neuroinflammation. 2016;
177. Ransohoff RM. A polarizing question: Do M1 and M2 microglia exist. Nature Neuroscience. 2016.
178. Cowan M, Petri WA. Microglia: Immune regulators of neurodevelopment. Frontiers in Immunology. 2018.
179. Sominsky L, De Luca S, Spencer SJ. Microglia: Key players in neurodevelopment and neuronal plasticity. International Journal of Biochemistry and Cell Biology. 2018.
180. Fourgeaud L, Traves PG, Tufail Y, Leal-Bailey H, Lew ED, Burrola PG, et al. TAM receptors regulate multiple features of microglial physiology. Nature. 2016;
181. Marín-Teva JL, Dusart I, Colin C, Gervais A, Van Rooijen N, Mallat M. Microglia Promote the Death of Developing Purkinje Cells. Neuron. 2004;
182. Frade JM, Barde YA. Microglia-derived nerve growth factor causes cell death in the developing retina. Neuron. 1998;
183. Sedel F. Macrophage-Derived Tumor Necrosis Factor , an Early Developmental Signal for Motoneuron Death. J Neurosci. 2004;
184. Cunningham CL, Martinez-Cerdeno V, Noctor SC. Microglia Regulate the Number of Neural Precursor Cells in the Developing Cerebral Cortex. J Neurosci. 2013;
185. Sierra A, Encinas JM, Deudero JJP, Chancey JH, Enikolopov G, Overstreet-Wadiche LS, et al. Microglia shape adult hippocampal neurogenesis through apoptosis-coupled phagocytosis. Cell Stem Cell. 2010;
186. Stevens B, Allen NJ, Vazquez LE, Howell GR, Christopherson KS,

Nouri N, et al. The Classical Complement Cascade Mediates CNS Synapse Elimination. Cell. 2007;

187. Chu Y, Jin X, Parada I, Pesic A, Stevens B, Barres B, et al. Enhanced synaptic connectivity and epilepsy in C1q knockout mice. Proc Natl Acad Sci. 2010;
188. Paolicelli RC, Bolasco G, Pagani F, Maggi L, Scianni M, Panzanelli P, et al. Synaptic pruning by microglia is necessary for normal brain development. Science (80-). 2011;
189. Schafer DP, Lehrman EK, Kautzman AG, Koyama R, Mardinly AR, Yamasaki R, et al. Microglia Sculpt Postnatal Neural Circuits in an Activity and Complement-Dependent Manner. Neuron. 2012;
190. Mody M, Cao Y, Cui Z, Tay KY, Shyong A, Shimizu E, et al. Genome-wide gene expression profiles of the developing mouse hippocampus. Proc Natl Acad Sci U S A. 2001;
191. Cardona AE, Pioro EP, Sasse ME, Kostenko V, Cardona SM, Dijkstra IM, et al. Control of microglial neurotoxicity by the fractalkine receptor. Nat Neurosci. 2006;
192. Liang KJ, Lee JE, Wang YD, Ma W, Fontainhas AM, Fariss RN, et al. Regulation of dynamic behavior of retinal microglia by CX3CR1 signaling. Investig Ophthalmol Vis Sci. 2009;
193. Ruitenberg MJ, Vukovic J, Blomster L, Hall JM, Jung S, Filgueira L, et al. CX3CL1/fractalkine regulates branching and migration of monocyte-derived cells in the mouse olfactory epithelium. J Neuroimmunol. 2008;
194. Harrison JK, Jiang Y, Chen S, Xia Y, Maciejewski D, McNamara RK, et al. Role for neuronally derived fractalkine in mediating interactions between neurons and CX3CR1-expressing microglia. Proc Natl Acad Sci. 1998;
195. Jung S, Aliberti J, Graemmel P, Sunshine MJ, Kreutzberg GW, Sher A, et al. Analysis of Fractalkine Receptor CX3CR1 Function by Targeted Deletion and Green Fluorescent Protein Reporter Gene Insertion. Mol Cell Biol. 2000;
196. Vainchtein ID, Chin G, Cho FS, Kelley KW, Miller JG, Chien EC, et al. Astrocyte-derived interleukin-33 promotes microglial synapse engulfment and neural circuit development. Science (80-). 2018;

197. Filipello F, Morini R, Corradini I, Zerbi V, Canzi A, Michalski B, et al. The Microglial Innate Immune Receptor TREM2 Is Required for Synapse Elimination and Normal Brain Connectivity. Immunity. 2018;
198. Parkhurst CN, Yang G, Ninan I, Savas JN, Yates JR, Lafaille JJ, et al. Microglia promote learning-dependent synapse formation through brain-derived neurotrophic factor. Cell. 2013;155(7):1596–609.
199. Li Y, Du X, Liu C, Wen Z, Du J. Reciprocal regulation between resting microglial dynamics and neuronal activity in vivo. Dev Cell. 2012;
200. Dissing-Olesen L, LeDue JM, Rungta RL, Hefendehl JK, Choi HB, MacVicar BA. Activation of Neuronal NMDA Receptors Triggers Transient ATP-Mediated Microglial Process Outgrowth. J Neurosci. 2014;
201. Lloyd AF, Miron VE. The pro-remyelination properties of microglia in the central nervous system. Nature Reviews Neurology. 2019.
202. Salter MW, Stevens B. Microglia emerge as central players in brain disease. Nature Medicine. 2017.
203. Serrano-Pozo A, Frosch MP, Masliah E, Hyman BT. Neuropathological alterations in Alzheimer disease. Cold Spring Harb Perspect Med. 2011;
204. Bachiller S, Jiménez-Ferrer I, Paulus A, Yang Y, Swanberg M, Deierborg T, et al. Microglia in Neurological Diseases: A Road Map to Brain-Disease Dependent-Inflammatory Response. Front Cell Neurosci. 2018;
205. McGeer PL, Itagaki S, Tago H, McGeer EG. Reactive microglia in patients with senile dementia of the Alzheimer type are positive for the histocompatibility glycoprotein HLA-DR. Neurosci Lett. 1987;
206. Doens D, Fernández PL. Microglia receptors and their implications in the response to amyloid β for Alzheimer's disease pathogenesis. Journal of Neuroinflammation. 2014.
207. Frick LR, Williams K, Pittenger C. Microglial dysregulation in psychiatric disease. Clin Dev Immunol [Internet]. 2013/04/18. 2013;2013:608654. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23690824
208. Huang KL, Marcora E, Pimenova AA, Di Narzo AF, Kapoor M, Jin SC,

et al. A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease. Nat Neurosci. 2017;
209. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell. 2013;
210. Feinberg I. Schizophrenia: Caused by a fault in programmed synaptic elimination during adolescence? J Psychiatr Res. 1982;
211. Sekar A, Bialas AR, De Rivera H, Davis A, Hammond TR, Kamitaki N, et al. Schizophrenia risk from complex variation of complement component 4. Nature. 2016;
212. Bialas AR, Stevens B. TGF-β signaling regulates neuronal C1q expression and developmental synaptic refinement. Nat Neurosci. 2013;
213. Sellgren CM, Gracias J, Watmuff B, Biag JD, Thanos JM, Whittredge PB, et al. Increased synapse elimination by microglia in schizophrenia patient-derived models of synaptic pruning. Nat Neurosci. 2019;
214. Glausier JR, Lewis DA. Dendritic spine pathology in schizophrenia. Neuroscience. 2013.
215. Kobayashi K, Imagama S, Ohgomori T, Hirano K, Uchimura K, Sakamoto K, et al. Minocycline selectively inhibits M1 polarization of microglia. Cell Death Dis. 2013;
216. Oya K, Kishi T, Iwata N. Efficacy and tolerability of minocycline augmentation therapy in schizophrenia: a systematic review and meta-analysis of randomized controlled trials. Hum Psychopharmacol Clin Exp [Internet]. 2014 Sep 1;29(5):483–91. Available from: https://doi.org/10.1002/hup.2426
217. Solmi M, Veronese N, Thapa N, Facchini S, Stubbs B, Fornaro M, et al. Systematic review and meta-analysis of the efficacy and safety of minocycline in schizophrenia. CNS Spectr [Internet]. 2017/02/09. 2017;22(5):415–26. Available from: https://www.cambridge.org/core/article/systematic-review-and-metaanalysis-of-the-efficacy-and-safety-of-minocycline-in-schizophrenia/2CA00F67A176134F280DFD4380001A13
218. Levy SE, Mandell DS, Schultz RT. ALevy, S. E., Mandell, D. S., &

Schultz, R. T. (2009). Autism. Lancet, 374, 1627–1638. doi:10.1016/S0140-6736(09)61376-3utism. Lancet. 2009;
219. Shaw C, Sheth S, Li D, Tomljenovic L. Etiology of autism spectrum disorders: Genes, environment, or both? OA Autism. 2014;
220. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet. 2014;
221. Poustka F. The neurobiology of autism. In: Autism and Pervasive Developmental Disorders, Second Edition. 2007.
222. Gupta S, Ellis SE, Ashar FN, Moes A, Bader JS, Zhan J, et al. Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. Nat Commun. 2014;
223. Ishizuka K, Fujita Y, Kawabata T, Kimura H, Iwayama Y, Inada T, et al. Rare genetic variants in CX3CR1 and their contribution to the increased risk of schizophrenia and autism spectrum disorders. Transl Psychiatry. 2017;
224. Rogers JT, Morganti JM, Bachstetter AD, Hudson CE, Peters MM, Grimmig BA, et al. CX3CR1 Deficiency Leads to Impairment of Hippocampal Cognitive Function and Synaptic Plasticity. J Neurosci. 2011;
225. Zhan Y, Paolicelli RC, Sforazzini F, Weinhard L, Bolasco G, Pagani F, et al. Deficient neuron-microglia signaling results in impaired functional brain connectivity and social behavior. Nat Neurosci. 2014;
226. Hoshiko M, Arnoux I, Avignone E, Yamamoto N, Audinat E. Deficiency of the Microglial Receptor CX3CR1 Impairs Postnatal Functional Development of Thalamocortical Synapses in the Barrel Cortex. J Neurosci. 2012;
227. Paolicelli RC, Bisht K, Tremblay M-Ã. Fractalkine regulation of microglial physiology and consequences on the brain and behavior. Front Cell Neurosci. 2014;
228. Tansey KE, Hill MJ. Enrichment of schizophrenia heritability in both neuronal and glia cell regulatory elements. Transl Psychiatry. 2018;
229. Sims R, van der Lee SJ, Naj AC, Bellenguez C, Badarinarayan N,

Jakobsdottir J, et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. Nat Genet. 2017;

230. Patterson PH. Immune involvement in schizophrenia and autism: etiology, pathology and animal models. Behav Brain Res. 2009;
231. Brown AS, Derkits EJ. Prenatal infection and schizophrenia: A review of epidemiologic and translational studies. American Journal of Psychiatry. 2010.
232. Knuesel I, Chicha L, Britschgi M, Schobel SA, Bodmer M, Hellings JA, et al. Maternal immune activation and abnormal brain development across CNS disorders. Nature Reviews Neurology. 2014.
233. Instanes JT, Halmøy A, Engeland A, Haavik J, Furu K, Klungsøyr K. Attention-Deficit/Hyperactivity Disorder in Offspring of Mothers With Inflammatory and Immune System Diseases. Biol Psychiatry. 2017;
234. Hornig M, Bresnahan MA, Che X, Schultz AF, Ukaigwe JE, Eddy ML, et al. Prenatal fever and autism risk. Mol Psychiatry. 2018;
235. Tansey KE, Cameron D, Hill MJ. Genetic risk for Alzheimer's disease is concentrated in specific macrophage and microglial transcriptional networks. Genome Med. 2018;
236. Gosselin D, Skola D, Coufal NG, Holtman IR, Schlachetzki JCM, Sajti E, et al. An environment-dependent transcriptional network specifies human microglia identity. Science (80-) [Internet]. 2017 [cited 2017 Jun 16]; Available from: http://science.sciencemag.org/content/early/2017/05/24/science.aal3222
237. Andrews S, Babraham Bioinformatics. FastQC: A quality control tool for high throughput sequence data. Manual. 2010.
238. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;
239. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;
240. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment / Map format and SAMtools. Bioinformatics.

2009;
241. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;
242. Stark R, Brown G. DiffBind : differential binding analysis of ChIP-Seq peak data. Cancer Res. 2011;
243. Benner C, Heinz S, Glass CK. HOMER - Software for motif discovery and next generation sequencing analysis. Http://Homer.Ucsd.Edu/. 2017.
244. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-RR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet [Internet]. 2015;47(11):1228–35. Available from: http://dx.doi.org/10.1038/ng.3404%5Cnhttp://10.1038/ng.3404%5Cnhttp://www.nature.com/ng/journal/v47/n11/abs/ng.3404.html#supplementary-information
245. Kierdorf K, Prinz M. Factors regulating microglia activation. Front Cell Neurosci. 2013;
246. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat Genet. 2019;
247. Luciano M, Hagenaars SP, Davies G, Hill WD, Clarke TK, Shirali M, et al. Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. Nat Genet. 2018;
248. Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet. 2018;
249. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;
250. Huang K, Marcora E, Pimenova A, Di Narzo A, Kapoor M, Jin SC, et al. A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease. bioRxiv. 2017;
251. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, Replogle JM, et al.

Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science (80-). 2014;

252. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. 2016;
253. Smith AM, Gibbons HM, Oldfield RL, Bergin PM, Mee EW, Faull RLM, et al. The transcription factor PU.1 is critical for viability and function of human brain microglia. Glia. 2013;
254. Lannes N, Eppler E, Etemad S, Yotovski P, Filgueira L. Microglia at center stage: A comprehensive review about the versatile and unique residential macrophages of the central nervous system. Oncotarget. 2017.
255. Ghisletti S, Barozzi I, Mietton F, Polletti S, De Santa F, Venturini E, et al. Identification and Characterization of Enhancers Controlling the Inflammatory Gene Expression Program in Macrophages. Immunity. 2010;
256. Zaret KS, Carroll JS. Pioneer transcription factors: Establishing competence for gene expression. Genes and Development. 2011.
257. Olmos-Alonso A, Schetters STT, Sri S, Askew K, Mancuso R, Vargas-Caballero M, et al. Pharmacological targeting of CSF1R inhibits microglial proliferation and prevents the progression of Alzheimer's-like pathology. Brain. 2016;
258. Tay TL, Béchade C, D'Andrea I, St-Pierre MK, Henry MS, Roumier A, et al. Microglia gone rogue: Impacts on psychiatric disorders across the lifespan. Frontiers in Molecular Neuroscience. 2018.
259. Mondelli V, Vernon AC, Turkheimer F, Dazzan P, Pariante CM. Brain microglia in psychiatric disorders. The Lancet Psychiatry. 2017.
260. Stan AD, Ghose S, Gao XM, Roberts RC, Lewis-Amezcua K, Hatanpaa KJ, et al. Human postmortem tissue: What quality markers matter? Brain Res. 2006;
261. Ni G, Moser G, Ripke S, Neale BM, Corvin A, Walters JTR, et al. Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood. Am J Hum Genet. 2018;

262. Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. Identification of transcription factor binding sites using ATAC-seq. Genome Biol [Internet]. 2019;20(1):45. Available from: https://doi.org/10.1186/s13059-019-1642-2
263. Soskic B, Cano-Gamez E, Smyth DJ, Rowan WC, Nakic N, Esparza-Gordillo J, et al. Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. bioRxiv. 2019;
264. Mizee MR, Miedema SSM, van der Poel M, Adelia, Schuurman KG, van Strien ME, et al. Isolation of primary microglia from the human post-mortem brain: effects of ante- and post-mortem variables. Acta Neuropathol Commun. 2017;
265. Butovsky O, Jedrychowski MP, Moore CS, Cialic R, Lanser AJ, Gabriely G, et al. Identification of a unique TGF-β–dependent molecular and functional signature in microglia. Nat Neurosci. 2013;17(1).
266. Bennett FC, Bennett ML, Yaqoob F, Mulinyawe SB, Grant GA, Hayden Gephart M, et al. A Combination of Ontogeny and CNS Environment Establishes Microglial Identity. Neuron. 2018;
267. Greter M, Lelios I, Pelczar P, Hoeffel G, Price J, Leboeuf M, et al. Stroma-Derived Interleukin-34 Controls the Development and Maintenance of Langerhans Cells and the Maintenance of Microglia. Immunity. 2012;
268. Haenseler W, Sansom SN, Buchrieser J, Newey SE, Moore CS, Nicholls FJ, et al. A Highly Efficient Human Pluripotent Stem Cell Microglia Model Displays a Neuronal-Co-culture-Specific Expression Profile and Inflammatory Response. Stem Cell Reports. 2017;8(6).
269. Griciuc A, Serrano-Pozo A, Parrado AR, Lesinski AN, Asselin CN, Mullin K, et al. Alzheimer's disease risk gene cd33 inhibits microglial uptake of amyloid beta. Neuron. 2013;
270. Timmerman R, Burm SM, Bajramovic JJ. An overview of in vitro methods to study microglia. Frontiers in Cellular Neuroscience. 2018.
271. Blasi E, Barluzzi R, Bocchini V, Mazzolla R, Bistoni F. Immortalization of murine microglial cells by a v-raf / v-myc carrying retrovirus. J Neuroimmunol. 1990;

272. Smith AM, Dragunow M. The human side of microglia. Trends in Neurosciences. 2014.
273. Stansley B, Post J, Hensley K. A comparative review of cell culture systems for the study of microglial biology in Alzheimer's disease. J Neuroinflammation. 2012;
274. Das A, Kim SH, Arifuzzaman S, Yoon T, Chai JC, Lee YS, et al. Transcriptome sequencing reveals that LPS-triggered transcriptional responses in established microglia BV2 cell lines are poorly representative of primary microglia. J Neuroinflammation. 2016;
275. Melief J, Sneeboer MAM, Litjens M, Ormel PR, Palmen SJMC, Huitinga I, et al. Characterizing primary human microglia: A comparative study with myeloid subsets and culture models. Glia. 2016;
276. Omole AE, Fakoya AOJ. Ten years of progress and promise of induced pluripotent stem cells: Historical origins, characteristics, mechanisms, limitations, and potential applications. PeerJ. 2018;
277. Abud EM, Ramirez RN, Martinez ES, Healy LM, Nguyen CHH, Newman SA, et al. iPSC-Derived Human Microglia-like Cells to Study Neurological Diseases. Neuron. 2017;94(2):278-293.e9.
278. Pandya H, Shen MJ, Ichikawa DM, Sedlock AB, Choi Y, Johnson KR, et al. Differentiation of human and murine induced pluripotent stem cells to microglia-like cells. Nat Neurosci [Internet]. 2017;20(5):753–9. Available from: http://www.nature.com/doifinder/10.1038/nn.4534
279. Muffat J, Li Y, Yuan B, Mitalipova M, Omer A, Corcoran S, et al. Efficient derivation of microglia-like cells from human pluripotent stem cells. Nat Med [Internet]. 2016;22(11):1358–67. Available from: http://dx.doi.org/10.1038/nm.4189%5Cnhttp://10.1038/nm.4189%5Cnhttp://www.nature.com/nm/journal/v22/n11/abs/nm.4189.html#supplementary-information
280. Gosselin D, Skola D, Coufal NG, Holtman IR, Schlachetzki JCM, Sajti E, et al. An environment-dependent transcriptional network specifies human microglia identity. Science (80-). 2017;
281. Janabi N, Peudenier S, Héron B, Ng KH, Tardieu M. Establishment of human microglial cell lines after transfection of primary cultures of

embryonic microglial cells with the SV40 large T antigen. Neurosci Lett. 1995;
282. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods. 2017;
283. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol [Internet]. 2011;29(1):24–6. Available from: https://doi.org/10.1038/nbt.1754
284. Groth D, Hartmann S, Klie S, Selbig J. Principal components analysis. Methods Mol Biol. 2013;
285. Pundhir S, Bratt Lauridsen FK, Schuster MB, Jakobsen JS, Ge Y, Schoof EM, et al. Enhancer and Transcription Factor Dynamics during Myeloid Differentiation Reveal an Early Differentiation Block in Cebpa null Progenitors. Cell Rep. 2018;
286. Zhang X, Wu J, Luo S, Lechler T, Zhang JY. FRA1 promotes squamous cell carcinoma growth and metastasis through distinct AKT and c-Jun dependent mechanisms. Oncotarget. 2016;
287. Kim S, Yu NK, Kaang BK. CTCF as a multifunctional protein in genome regulation and gene expression. Experimental & molecular medicine. 2015.
288. Klenova EM, Morse HC, Ohlsson R, Lobanenkov V V. The novel BORIS + CTCF gene family is uniquely involved in the epigenetics of normal biology and cancer. Semin Cancer Biol. 2002;
289. Pugacheva EM, Suzuki T, Pack SD, Kosaka-Suzuki N, Yoon J, Vostrov AA, et al. The Structural Complexity of the Human BORIS Gene in Gametogenesis and Cancer. PLoS One. 2010;
290. Hoek RH, Ruuls SR, Murphy CA, Wright GJ, Goddard R, Zurawski SM, et al. Down-regulation of the macrophage lineage through interaction with OX2 (CD200). Science (80-). 2000;
291. Baghdadi M, Umeyama Y, Hama N, Kobayashi T, Han N, Wada H, et al. Interleukin-34, a comprehensive review. Journal of Leukocyte Biology. 2018.
292. Vainchtein ID, Chin G, Cho FS, Kelley KW, Miller JG, Chien EC, et al.

Astrocyte-derived interleukin-33 promotes microglial synapse engulfment and neural circuit development. Science (80-). 2018;
293. Norden DM, Fenn AM, Dugan A, Godbout JP. TGFβ produced by IL-10 redirected astrocytes attenuates microglial activation. Glia. 2014;
294. Schilling T, Nitsch R, Heinemann U, Haas D, Eder C. Astrocyte-released cytokines induce ramification and outward K+ channel expression in microglia via distinct signalling pathways. Eur J Neurosci. 2001;
295. Douvaras P, Sun B, Wang M, Kruglikov I, Lallos G, Zimmer M, et al. Directed Differentiation of Human Pluripotent Stem Cells to Microglia. Stem Cell Reports. 2017;
296. Lancaster MA, Renner M, Martin CA, Wenzel D, Bicknell LS, Hurles ME, et al. Cerebral organoids model human brain development and microcephaly. Nature. 2013;
297. Karzbrun E, Reiner O. Brain organoids—A bottom-up approach for studying human neurodevelopment. Bioengineering. 2019.
298. Yakoub AM. Cerebral organoids exhibit mature neurons and astrocytes and recapitulate electrophysiological activity of the human brain. Neural Regeneration Research. 2019.
299. Ormel PR, Vieira de Sá R, van Bodegraven EJ, Karst H, Harschnitz O, Sneeboer MAM, et al. Microglia innately develop within cerebral organoids. Nat Commun. 2018;
300. Mariani J, Coppola G, Zhang P, Abyzov A, Provini L, Tomasini L, et al. FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. Cell. 2015;
301. Rubenstein JLR. Three hypotheses for developmental defects that may underlie some forms of autism spectrum disorder. Current Opinion in Neurology. 2010.
302. Murray RM, Lewis SW. Is schizophrenia a neurodevelopmental disorder? British Medical Journal (Clinical research ed.). 1988.
303. Owen MJ, O'Donovan MC, Thapar A, Craddock N. Neurodevelopmental hypothesis of schizophrenia. Br J Psychiatry. 2011;
304. Moran P, Stokes J, Marr J, Bock G, Desbonnet L, Waddington J, et al.

Gene × Environment Interactions in Schizophrenia: Evidence from Genetic Mouse Models. Neural Plast. 2016;
305. Menassa DA, Gomez-Nicola D. Microglial dynamics during human brain development. Frontiers in Immunology. 2018.
306. Lenz KM, Nelson LH. Microglia and beyond: Innate immune cells as regulators of brain development and behavioral function. Frontiers in Immunology. 2018.
307. O'Brien HE, Hannon E, Hill MJ, Toste CC, Robertson MJ, Morgan JE, et al. Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. Genome Biol. 2018;
308. de la Torre-Ubieta L, Stein JL, Won H, Opland CK, Liang D, Lu D, et al. The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. Cell. 2018;
309. Cossarizza A, Chang HD, Radbruch A, Akdis M, Andrä I, Annunziato F, et al. Guidelines for the use of flow cytometry and cell sorting in immunological studies. Eur J Immunol. 2017;
310. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;
311. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;
312. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;
313. Iotchkova V, Ritchie GRS, Geihs M, Morganella S, Min JL, Walter K, et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. Nat Genet. 2019;
314. The UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;
315. Smith AM, Gibbons HM, Oldfield RL, Bergin PM, Mee EW, Faull RLM, et al. The transcription factor PU.1 is critical for viability and function of

human brain microglia. Glia. 2013;61(6):929–42.
316. Nagy C, Maheu M, Lopez JP, Vaillancourt K, Cruceanu C, Gross JA, et al. Effects of Postmortem Interval on Biomolecule Integrity in the Brain. J Neuropathol Exp Neurol. 2015;
317. Jarmasz JS, Stirton H, Davie JR, Del Bigio MR. DNA methylation and histone post-translational modification stability in post-mortem brain tissue. Clin Epigenetics [Internet]. 2019;11(1):5. Available from: https://doi.org/10.1186/s13148-018-0596-7
318. Lopez-Atalaya JP, Askew KE, Sierra A, Gomez-Nicola D. Development and maintenance of the brain's immune toolkit: Microglia and non-parenchymal brain macrophages. Developmental Neurobiology. 2018.
319. Faraco G, Park L, Anrather J, Iadecola C. Brain perivascular macrophages: characterization and functional roles in health and disease. Journal of Molecular Medicine. 2017.
320. Cao M, Wang Z, He Y. Connectomics in psychiatric research: Advances and applications. Neuropsychiatric Disease and Treatment. 2015.
321. Tønnesen S, Kaufmann T, Richard G, Doan NT, Alnæs D, van der Meer D, et al. Brain age prediction reveals aberrant brain white matter in schizophrenia and bipolar disorder: A multi-sample diffusion tensor imaging study. bioRxiv [Internet]. 2019 Jan 1;607754. Available from: http://biorxiv.org/content/early/2019/04/12/607754.abstract
322. F.K. J, A. K. Early life stress perturbs the function of microglia in the developing rodent brain: New insights and future challenges. Brain Behav Immun. 2018;
323. Delpech JC, Wei L, Hao J, Yu X, Madore C, Butovsky O, et al. Early life stress perturbs the maturation of microglia in the developing hippocampus. Brain Behav Immun. 2016;
324. Mattei D, Ivanov A, Ferrai C, Jordan P, Guneykaya D, Buonfiglioli A, et al. Maternal immune activation results in complex microglial transcriptome signature in the adult offspring that is reversed by minocycline treatment. Transl Psychiatry. 2017;
325. Bryois J, Garrett ME, Song L, Safi A, Giusti-Rodriguez P, Johnson GD, et al. Evaluation of chromatin accessibility in prefrontal cortex of

individuals with schizophrenia. Nat Commun. 2018;
326. Herz J, Filiano AJ, Smith A, Yogev N, Kipnis J. Myeloid Cells in the Central Nervous System. Immunity. 2017.
327. Li Q, Cheng Z, Zhou L, Darmanis S, Neff NF, Okamoto J, et al. Developmental Heterogeneity of Microglia and Brain Myeloid Cells Revealed by Deep Single-Cell RNA Sequencing. Neuron. 2019;
328. Polioudakis D, de la Torre-Ubieta L, Langerman J, Elkins AG, Shi X, Stein JL, et al. A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. Neuron. 2019;
329. Doorn KJ, BrevÃ© JJP, Drukarch B, Boddeke HW, Huitinga I, Lucassen PJ, et al. Brain region-specific gene expression profiles in freshly isolated rat microglia. Front Cell Neurosci. 2015;
330. Lavin Y, Winter D, Blecher-Gonen R, David E, Keren-Shaul H, Merad M, et al. Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. Cell. 2014;159(6).
331. Fan X, Dong J, Zhong S, Wei Y, Wu Q, Yan L, et al. Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis. Cell Res. 2018;
332. Iotchkova V, Ritchie GRS, Geihs M, Morganella S, Min JL, Walter K, et al. GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. bioRxiv [Internet]. 2016 Jan 1;85738. Available from: http://biorxiv.org/content/early/2016/11/07/085738.abstract
333. Chen X, Miragaia RJ, Natarajan KN, Teichmann SA. A rapid and robust method for single cell chromatin accessibility profiling. Nat Commun. 2018;
334. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22(9):1748–59.
335. Deplancke B, Alpern D, Gardeux V. The Genetics of Transcription Factor DNA Binding Variation. Cell. 2016.
336. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;

337. Coetzee SG, Coetzee GA, Hazelett DJ. MotifbreakR: An R/Bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics. 2015;
338. Coetzee SG, Pierce S, Brundin P, Brundin L, Hazelett DJ, Coetzee GA. Enrichment of risk SNPs in regulatory regions implicate diverse tissues in Parkinson's disease etiology. Sci Rep. 2016;
339. Nguyen NTT, Contreras-Moreira B, Castro-Mondragon JA, Santana-Garcia W, Ossio R, Robles-Espinoza CD, et al. RSAT 2018: Regulatory sequence analysis tools 20th anniversary. Nucleic Acids Res. 2018;
340. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;
341. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. 2014;
342. Wang J, Zhuang J, Iyer S, Lin XY, Whitfield TW, Greven MC, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 2012;
343. Kulakovskiy I V., Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2018;
344. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, et al. Evaluation of methods for modeling transcription factor sequence specificity. Nat Biotechnol. 2013;
345. Ward LD, Kellis M. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;
346. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, et al. Cistrome Data Browser: A data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res. 2017;
347. Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, et al. Cistrome Data Browser: Expanded datasets and new tools for gene regulatory

analysis. Nucleic Acids Res. 2019;
348. Gates LA, Foulds CE, O'Malley BW. Histone Marks in the 'Driver's Seat': Functional Roles in Steering the Transcription Cycle. Trends in Biochemical Sciences. 2017.
349. Pham TH, Benner C, Lichtinger M, Schwarzfischer L, Hu Y, Andreesen R, et al. Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. Blood. 2012;
350. Wei GH, Badis G, Berger MF, Kivioja T, Palin K, Enge M, et al. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J. 2010;
351. Sharrocks AD. The ETS-domain transcription factor family. Nat Rev Mol Cell Biol. 2001;
352. Hollenhorst PC, McIntosh LP, Graves BJ. Genomic and Biochemical Insights into the Specificity of ETS Transcription Factors. Annu Rev Biochem. 2011;
353. Nerlov C. The C/EBP family of transcription factors: a paradigm for interaction between gene expression and proliferation control. Trends in Cell Biology. 2007.
354. Ejarque-Ortiz A, Medina MG, Tusell JM, Pérez-González AP, Serratosa J, Saura J. Upregulation of CCAAT/enhancer binding protein β in activated astrocytes and microglia. Glia. 2007;
355. Strohmeyer R, Shelton J, Lougheed C, Breitkopf T. CCAAT-enhancer binding protein-β expression and elevation in Alzheimer's disease and microglial cell cultures. PLoS One. 2014;
356. Wang ZH, Gong K, Liu X, Zhang Z, Sun X, Wei ZZ, et al. C/EBPβ regulates delta-secretase expression and mediates pathogenesis in mouse models of Alzheimer's disease. Nat Commun. 2018;
357. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. The role of DNA shape in protein-DNA recognition. Nature. 2009;
358. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of Specificity in Protein-DNA Recognition. Annu Rev Biochem. 2010;
359. Parker SCJ, Hansen L, Abaan HO, Tullius TD, Margulies EH. Local DNA topography correlates with functional noncoding regions of the

human genome. Science (80-). 2009;
360. Hellman LM, Fried MG. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nat Protoc. 2007;
361. Holden NS, Tacon CE. Principles and problems of the electrophoretic mobility shift assay. J Pharmacol Toxicol Methods. 2011;
362. Miller DE, Patel ZH, Lu X, Lynch AT, Weirauch MT, Kottyan LC. Screening for Functional Non-coding Genetic Variants Using Electrophoretic Mobility Shift Assay (EMSA) and DNA-affinity Precipitation Assay (DAPA). J Vis Exp. 2016;
363. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science (80-). 2013;
364. Eid A, Alshareef S, Mahfouz MM. CRISPR base editors: genome editing without double-stranded breaks. Biochem J. 2018;
365. Schmitt AD, Hu M, Ren B. Genome-wide mapping and analysis of chromosome architecture. Nature Reviews Molecular Cell Biology. 2016.
366. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017;
367. Nott A, Holtman IR, Coufal NG, Schlachetzki JCM, Yu M, Hu R, et al. Cell type-specific enhancer-promoter connectivity maps in the human brain and disease risk association. bioRxiv [Internet]. 2019 Jan 1;778183. Available from: http://biorxiv.org/content/early/2019/09/22/778183.abstract
368. Novikova G, Kapoor M, TCW J, Abud EM, Efthymiou AG, Cheng H, et al. Integration of Alzheimer's disease genetics and myeloid genomics reveals novel disease risk mechanisms. bioRxiv [Internet]. 2019 Jan 1;694281. Available from: http://biorxiv.org/content/early/2019/08/12/694281.abstract
369. Kierdorf K, Prinz M. Factors regulating microglia activation. Frontiers in Cellular Neuroscience. 2013.
370. Heinz S, Romanoski CE, Benner C, Allison KA, Kaikkonen MU, Orozco LD, et al. Effect of natural genetic variation on enhancer selection and function. Nature. 2013;
371. Cardno AG, Owen MJ. Genetic relationships between schizophrenia,

bipolar disorder, and schizoaffective disorder. Schizophr Bull. 2014;
372. Clementz BA, Sweeney JA, Hamm JP, Ivleva EI, Ethridge LE, Pearlson GD, et al. Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers. Am J Psychiatry [Internet]. 2015 Dec 7;173(4):373–84. Available from: https://doi.org/10.1176/appi.ajp.2015.14091200
373. Forstner AJ, Hecker J, Hofmann A, Maaser A, Reinbold CS, Mühleisen TW, et al. Identification of shared risk loci and pathways for bipolar disorder and schizophrenia. PLoS One. 2017;
374. Ivleva E, Thaker G, Tamminga CA. Comparing genes and phenomenology in the major psychoses: Schizophrenia and bipolar 1 disorder. Schizophrenia Bulletin. 2008.
375. Schulze TG, Akula N, Breuer R, Steele J, Nalls MA, Singleton AB, et al. Molecular genetic overlap in bipolar disorder, schizophrenia, and major depressive disorder. World J Biol Psychiatry. 2014;
376. Owen MJ, O'Donovan MC, Thapar A, Craddock N. Neurodevelopmental hypothesis of schizophrenia. Br J Psychiatry. 2011;
377. Keshavan MS. Development, disease and degeneration in schizophrenia: A unitary pathophysiological model. J Psychiatr Res. 1999;
378. Keshavan MS, Hogarty GE. Brain maturational processes and delayed onset in schizophrenia. Development and Psychopathology. 1999.
379. Szepesi Z, Manouchehrian O, Bachiller S, Deierborg T. Bidirectional Microglia–Neuron Communication in Health and Disease. Frontiers in Cellular Neuroscience. 2018.
380. Eyo UB, Wu LJ. Bidirectional microglia-neuron communication in the healthy brain. Neural Plasticity. 2013.
381. Bohlen CJ, Bennett FC, Tucker AF, Collins HY, Mulinyawe SB, Barres BA. Diverse Requirements for Microglial Survival, Specification, and Function Revealed by Defined-Medium Cultures. Neuron. 2017;
382. Butovsky O, Jedrychowski MP, Moore CS, Cialic R, Lanser AJ, Gabriely G, et al. Identification of a unique TGF-β-dependent molecular and functional signature in microglia. Nat Neurosci. 2014;

383. Li Q, Barres BA. Microglia and macrophages in brain homeostasis and disease. Nature Reviews Immunology. 2018.
384. Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol. 2018;
385. Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol. 2014;
386. Lawson LJ, Perry VH, Dri P, Gordon S. Heterogeneity in the distribution and morphology of microglia in the normal adult mouse brain. Neuroscience. 1990;
387. Grabert K, Michoel T, Karavolos MH, Clohisey S, Kenneth Baillie J, Stevens MP, et al. Microglial brain regionâ 'dependent diversity and selective regional sensitivities to aging. Nat Neurosci. 2016;
388. de Haas AH, Boddeke HWGM, Biber K. Region-specific expression of immunoregulatory proteins on microglia in the healthy CNS. Glia. 2008;
389. Tay TL, Mai D, Dautzenberg J, Fernández-Klett F, Lin G, Sagar S, et al. A new fate mapping system reveals context-dependent random or clonal expansion of microglia. Nat Neurosci. 2017;
390. Olah M, Menon V, Habib N, Taga M, Yung C, Cimpean M, et al. A single cell-based atlas of human microglial states reveals associations with neurological disorders and histopathological features of the aging brain. bioRxiv [Internet]. 2018 Jan 1;343780. Available from: http://biorxiv.org/content/early/2018/06/11/343780.abstract
391. Guneykaya D, Ivanov A, Hernandez DP, Haage V, Wojtas B, Meyer N, et al. Transcriptional and Translational Differences of Microglia from Male and Female Brains. Cell Rep. 2018;
392. Kodama L, Gan L. Do Microglial Sex Differences Contribute to Sex Differences in Neurodegenerative Diseases? Trends Mol Med [Internet]. 2019;25(9):741–9. Available from: http://www.sciencedirect.com/science/article/pii/S1471491419301030
393. Mrdjen D, Fox EJ, Bukhari SA, Montine KS, Bendall SC, Montine TJ.

The basis of cellular and regional vulnerability in Alzheimer's disease. Acta Neuropathol [Internet]. 2019; Available from: https://doi.org/10.1007/s00401-019-02054-4

394. Holmes C, Cunningham C, Zotova E, Woolford J, Dean C, Kerr S, et al. Systemic inflammation and disease progression in alzheimer disease. Neurology. 2009;
395. Wendeln AC, Degenhardt K, Kaurani L, Gertig M, Ulas T, Jain G, et al. Innate immune memory in the brain shapes neurological disease hallmarks. Nature. 2018;556(7701).
396. Neher JJ, Cunningham C. Priming Microglia for Innate Immune Memory in the Brain. Trends in Immunology. 2019.
397. Holtman IR, Raj DD, Miller JA, Schaafsma W, Yin Z, Brouwer N, et al. Induction of a common microglia gene expression signature by aging and neurodegenerative conditions: a co-expression meta-analysis. Acta Neuropathol Commun. 2015;
398. Perry VH, Holmes C. Microglial priming in neurodegenerative disease. Nature Reviews Neurology. 2014.
399. Perry VH, Nicoll JAR, Holmes C. Microglia in neurodegenerative disease. Nature Reviews Neurology. 2010.
400. Casserly I, Topol E. Convergence of atherosclerosis and Alzheimer's disease: Inflammation, cholesterol, and misfolded proteins. Lancet. 2004.
401. Balakrishnan K, Verdile G, Mehta PD, Beilby J, Nolan D, Galvão DA, et al. Plasma Aβ42 correlates positively with increased body fat in healthy individuals. J Alzheimer's Dis. 2005;
402. Donath MY, Shoelson SE. Type 2 diabetes as an inflammatory disease. Nature Reviews Immunology. 2011.
403. Eikelenboom P, Hoozemans JJM, Veerhuis R, Van Exel E, Rozemuller AJM, Van Gool WA. Whether, when and how chronic inflammation increases the risk of developing late-onset Alzheimer's disease. Alzheimer's Research and Therapy. 2012.
404. de Wit E, de Laat W. A decade of 3C technologies: Insights into nuclear organization. Genes Dev. 2012;
405. Denker A, De Laat W. The second decade of 3C technologies:

Detailed insights into nuclear organization. Genes and Development. 2016.

8 Appendix

8.1 Garfield SNP enrichment tests

Table 8.1. Enrichment of brain disorder GWAS SNPs in conserved ENCODE esophageal bulk cell open chromatin regions

GWAS	GWAS Thresh	OR	p-value	cor. p-value	Beta	SE	CI95_L	CI95_U	No. Annot Thesh	N. Annot	N. Thresh	No. of SNPs
ADHD	1 x 10-5	1.02	0.975	1.000	0.018	0.593	-1.144	1.181	5	2,816	138	233,701
ADHD	1 x 10-8	7.67 x 10-8	0.998	1.000	-16.384	5872	-11525	11492	0	2,816	4	233,701
AUTISM	1 x 10-5	2.26	0.175	1.000	0.817	0.602	-0.363	1.998	3	3,330	103	357,689
AUTISM	1 x 10-8	1.07 x 10 -6	0.997	1.000	-13.745	4108	-8066	8039	0	3,330	1	357,689
BPD	1 x 10-5	1.35	0.415	1.000	0.299	0.367	-0.420	1.018	8	9,942	227	1,191,616
BPD	1 x 10-8	1.07 x 10-7	0.996	1.000	-16.052	2973	-5844	5812	0	9,942	8	1,191,616
LOAD	1 x 10-5	2.71	**0.014**	0.224	0.996	0.403	0.206	1.786	7	3,400	113	262,885
LOAD	1 x 10-8	1.81	0.423	1.000	0.592	0.739	-0.856	2.040	2	3,400	43	262,885
MDD	1 x 10-5	1.55	0.345	1.000	0.438	0.464	-0.471	1.347	5	9,708	159	1,157,454
MDD	1 x 10-8	2.51	0.392	1.000	0.921	1.077	-1.190	3.033	1	9,708	8	1,157,454
NEUROTICISM	1 x 10-5	1.18	0.445	1.000	0.169	0.221	-0.264	0.602	26	9,336	697	907,858
NEUROTICISM	1 x 10-8	0.65	0.554	1.000	-0.426	0.721	-1.839	0.986	3	9,336	84	907,858
SCZ	1 x 10-5	1.45	**0.031**	0.488	0.374	0.173	0.035	0.714	40	3,305	1,042	317,182
SCZ	1 x 10-8	1.73	0.085	1.000	0.547	0.318	-0.075	1.170	13	3,305	232	317,182
WGL	1 x 10-5	1.53	0.675	1.000	0.428	1.021	-1.574	2.429	1	6,197	68	891,538
WGL	1 x 10-8	7.20 x 10-7	0.997	1.000	-14.144	3630	-7129	7100	0	6,197	2	891,538

GWAS Thresh (GWAS significance threshold); OR (Odds ratio); p-value (emprical p-value of the significance of the observed enrichment); cor. p-value (p-value Bonferroni corrected for 16 tests); Beta (effect size from the general linear model, equal to log(OR)); SE (standard error from generalised linear model); CI95_L (lower boundary for the 95% CI of the effect size/beta); CI95_U (upper boundary for the 95% CI of the effect size/beta); No. Annot. Thresh (number of annotated variants within annotation passing GWAS thresh); N.Annot (number of variants within given annotation); N.Thresh (number of variants passing GWAS thresh); No. of SNPs (total number of LD pruned variants). ADHD (Attention deficit hyperactvity disorder); ASD (Autism spectrum disorder); BPD (Bipolar disorder); LOAD (Late onset Alzheimer's disorder); MDD (Major depressive disorder); SCZ (Schizophrenia); WGL (Wearer of glasses or contact lenses).

Table 8.2. Enrichment of brain disorder GWAS SNPs in conserved ENCODE colon bulk cell chromatin regions

GWAS	GWAS Thresh	OR	p-value	cor. p-value	Beta	SE	CI95_L	CI95_U	No. Annot Thesh	N. Annot	N. Thresh	No. of SNPs
ADHD	1 x 10-5	0.92	0.908	1.000	-0.083	0.719	-1.494	1.328	5	2,159	138	233,701
ADHD	1 x 10-8	8.13 x 10-8	0.998	1.000	-16.325	6786	-13318	13285	0	2,159	4	233,701
AUTISM	1 x 10-5	1.01	0.996	1.000	0.006	1.014	-1.983	1.994	1	2,461	103	357,689
AUTISM	1 x 10-8	7.86 x 10-7	0.998	1.000	-14.056	4800	-9422	9394	0	2,461	1	357,689
BPD	1 x 10-5	1.11	0.810	1.000	0.110	0.458	-0.787	1.007	5	7,417	227	1,191,616
BPD	1 x 10-8	1.11 x 10-7	0.996	1.000	-16.015	3476	-6829	6797	0	7,417	8	1,191,616
LOAD	1 x 10-5	1.94	0.202	1.000	0.663	0.519	-0.354	1.679	4	2,667	113	262,885
LOAD	1 x 10-8	6.40 x 10-8	0.995	1.000	-16.564	2516	-4949	4916	0	2,667	43	262,885
MDD	1 x 10-5	1.21	0.750	1.000	0.188	0.59	-0.969	1.344	3	7,241	159	1,157,454
MDD	1 x 10-8	9.54 x 10-8	0.996	1.000	-16.165	3445	-6768	6736	0	7,241	8	1,157,454
NEUROTICISM	1 x 10-5	1.08	0.779	1.000	0.074	0.265	-0.445	0.594	17	7,085	697	907,858
NEUROTICISM	1 x 10-8	1.87	0.230	1.000	0.624	0.519	-0.394	1.641	5	7,085	84	907,858
SCZ	1 x 10-5	1.90	**2.97 x 10-5**	**4.75 x 10-4**	0.644	0.178	0.295	0.994	40	2,527	1,042	317182
SCZ	1 x 10-8	3.16	**5.80 x 10-5**	**9.28 x 10-4**	1.150	0.286	0.589	1.710	16	2,527	232	317,182
WGL	1 x 10-5	3.71 x 10-6	0.978	1.000	-12.503	443	-881	856	0	4,595	68	891,538
WGL	1 x 10-8	6.32 x 10-7	0.997	1.000	-14.274	4261	-8366	8338	0	4,595	2	891,538

GWAS Thresh (GWAS significance threshold); OR (Odds ratio); p-value (emprical p-value of the significance of the observed enrichment); cor. p-value (p-value Bonferroni corrected for 16 tests); Beta (effect size from the general linear model, equal to log(OR)); SE (standard error from generalised linear model); CI95_L (lower boundary for the 95% CI of the effect size/beta); CI95_U (upper boundary for the 95% CI of the effect size/beta); No. Annot. Thresh (number of annotated variants within annotation passing GWAS thresh); N.Annot (number of variants within given annotation); N.Thresh (number of variants passing GWAS thresh); No. of SNPs (total number of LD pruned variants). ADHD (Attention deficit hyperactvity disorder); ASD (Autism spectrum disorder); BPD (Bipolar disorder); LOAD (Late onset Alzheimer's disorder); MDD (Major depressive disorder); SCZ (Schizophrenia); WGL (Wearer of glasses or contact lenses).

Table 8.3. Enrichment of brain disorder GWAS SNPs in conserved ENCODE heart bulk cell open chromatin regions

GWAS	GWAS Thresh	OR	p-value	cor. p-value	Beta	SE	CI95_L	CI95_U	No. Annot Thesh	N. Annot	N. Thresh	No. of SNPs
ADHD	1 x 10-5	2.58	0.189	1.000	0.948	0.721	-0.466	2.362	2	768	138	233,701
ADHD	1 x 10-8	8.50 x10-8	0.999	1.000	-16.280	11208	-21984	21952	0	768	4	233,701
AUTISM	1 x 10-5	2.8	0.310	1.000	1.031	1.015	-0.959	3.020	1	880	103	357,689
AUTISM	1 x 10-8	1.70 x 10-6	0.999	1.000	-13.284	7832	-15363	15337	0	880	1	357,689
BPD	1 x 10-5	0.63	0.639	1.000	-0.471	1.005	-2.441	1.499	1	2,653	227	1,191,616
BPD	1 x 10-8	1.18 x 10-7	0.998	1.000	-15.946	5700	-11187	11155	0	2,653	8	1,191,616
LOAD	1 x 10-5	2.83	0.150	1.000	1.040	0.722	-0.375	2.455	2	862	113	262,885
LOAD	1 x 10-8	3.52	0.220	1.000	1.257	1.024	-0.750	3.264	1	862	43	262,885
MDD	1 x 10-5	3.69 x 10-6	0.970	1.000	-12.509	329	-658	633	0	2,543	159	1,157,454
MDD	1 x 10-8	9.27 x 10-8	0.998	1.000	-16.193	5705	-11197.713	11165.326	0	2,543	8	1,157,454
NEUROTICISM	1 x 10-5	1.42	0.362	1.000	0.350	0.385	-0.404	1.104	9	2,449	697	907,858
NEUROTICISM	1 x 10-8	1.13 x 10-8	0.998	1.000	-18.303	6357	-12478	12441	0	2449	84	907,858
SCZ	1 x 10-5	1.67	**0.084**	1.000	0.514	0.298	-0.070	1.097	14	917	1,042	317,182
SCZ	1 x 10-8	3.31	**0.005**	0.076	1.197	0.427	0.367	2.027	7	917	232	317,182
WGL	1 x 10-5	1.00 x 10-5	0.980	1.000	-11.510	463	-918	895	0	1,540	68	891,538
WGL	1 x 10-8	6.47 x 10-7	0.998	1.000	-14.251	7347	-14413	14385	0	1,540	2	891,538

GWAS Thresh (GWAS significance threshold); OR (Odds ratio); p-value (emprical p-value of the significance of the observed enrichment); cor. p-value (p-value Bonferroni corrected for 16 tests); Beta (effect size from the general linear model, equal to log(OR)); SE (standard error from generalised linear model); CI95_L (lower boundary for the 95% CI of the effect size/beta); CI95_U (upper boundary for the 95% CI of the effect size/beta); No. Annot. Thresh (number of annotated variants within annotation passing GWAS thresh); N.Annot (number of variants within given annotation); N.Thresh (number of variants passing GWAS thresh); No. of SNPs (total number of LD pruned variants). ADHD (Attention deficit hyperactvity disorder); ASD (Autism spectrum disorder); BPD (Bipolar disorder); LOAD (Late onset Alzheimer's disorder); MDD (Major depressive disorder); SCZ (Schizophrenia); WGL (Wearer of glasses or contact lenses).

Table 8.4. Enrichment of brain disorder GWAS SNPs in conserved ENCODE liver bulk cell open chromatin regions

GWAS	GWAS Thresh	OR	p-value	cor. p-value	Beta	SE	CI95_L	CI95_U	No. Annot Thesh	N. Annot	N. Thresh	No. of SNPs
ADHD	1 x 10-5	5.48 x 10-8	0.996	1.000	-16.720	3664	-7198	7165	0	457	138	233,701
ADHD	1 x 10-8	7.54 x 10-8	0.999	1.000	-16.400	14616	-28665	28632	0	457	4	233,701
AUTISM	1 x 10-5	4.38	0.146	1.000	1.477	1.017	-0.516	3.470	1	553	103	357,689
AUTISM	1 x 10-8	1.55 x 10-6	0.999	1.000	-13.376	10050	-19713	19686	0	553	1	357,689
BPD	1 x 10-5	6.29 x 10-6	0.960	1.000	-11.976	236	-474	450	0	1,778	227	1,191,616
BPD	1 x 10-8	1.17 x 10-7	0.998	1.000	-15.959	7099	-13930	13898	0	1,778	8	1,191,616
LOAD	1 x 10-5	6.26	**0.002**	**0.035**	1.833	0.598	0.661	3.006	3	570	113	262,885
LOAD	1 x 10-8	5.93 x 10-8	0.998	1.000	-16.641	5440	-10680	10646	0	570	43	262,885
MDD	1 x 10-5	9.67 x 10-6	0.963	1.000	-11.543	2456	-493	470	0	1,730	159	1,157,454
MDD	1 x 10-8	8.72 x 10-8	0.998	1.000	-16.255	7060	-13853	13820	0	1,730	8	1,157,454
NEUROTICISM	1 x 10-5	1.12	0.821	1.000	0.114	0.506	-0.878	1.106	6	1,700	697	907,858
NEUROTICISM	1 x 10-8	1.78	0.569	1.000	0.576	1.011	-1.406	2.558	1	1,700	84	907,858
SCZ	1 x 10-5	1.57	0.245	1.000	0.450	0.387	-0.309	1.209	8	577	1,042	317,182
SCZ	1 x 10-8	2.63	0.101	1.000	0.968	0.590	-0.188	2.124	3	577	232	317,182
WGL	1 x 10-5	2.67 x 10-5	0.975	1.000	-10.530	336	-670	649	0	1,056	68	891,538
WGL	1 x 10-8	6.94 x 10-7	0.999	1.000	-14.181	8801	-17265	17237	0	1,056	2	891,538

GWAS Thresh (GWAS significance threshold); OR (Odds ratio); p-value (emprical p-value of the significance of the observed enrichment); cor. p-value (p-value Bonferroni corrected for 16 tests); Beta (effect size from the general linear model, equal to log(OR)); SE (standard error from generalised linear model); CI95_L (lower boundary for the 95% CI of the effect size/beta); CI95_U (upper boundary for the 95% CI of the effect size/beta); No. Annot. Thresh (number of annotated variants within annotation passing GWAS thresh); N.Annot (number of variants within given annotation); N.Thresh (number of variants passing GWAS thresh); No. of SNPs (total number of LD pruned variants). ADHD (Attention deficit hyperactvity disorder); ASD (Autism spectrum disorder); BPD (Bipolar disorder); LOAD (Late onset Alzheimer's disorder); MDD (Major depressive disorder); SCZ (Schizophrenia); WGL (Wearer of glasses or contact lenses).

Table 8.5. Enrichment of brain disorder GWAS SNPs in conserved ENCODE stomach bulk cell open chromatin regions

GWAS	GWAS Thresh	OR	p-value	cor. p-value	Beta	SE	CI95_L	CI95_U	No. Annot Thesh	N. Annot	N. Thresh	No. of SNPs
ADHD	1 x 10-5	5.10 x 10-8	0.999	1.000	-16.792	9928	-19476	19442	0	62	138	233,701
ADHD	1 x 10-8	6.27 x 10-8	1.000	1.000	-16.585	39128	-76709	76676	0	62	4	233,701
AUTISM	1 x 10-5	4.73 x10-5	0.982	1.000	-9.959	436	-865	845	0	88	103	357,689
AUTISM	1 x 10-8	0.387	1.000	1.000	-0.948	31137	-61030	61028	0	88	1	357,689
BPD	1 x 10-5	4.49 x 10-5	0.965	1.000	-10.011	225	-451	431	0	262	227	1,191,616
BPD	1 x 10-8	1.19 x 10-7	0.999	1.000	-15.947	18556	-36386	36354	0	262	8	1191616
LOAD	1 x 10-5	12.24	**0.014**	0.228	2.505	1.022	0.502	4.508	1	94	113	262,885
LOAD	1 x 10-8	5.48	0.999	1.000	-16.719	13653	-26776	26743	0	94	43	262,885
MDD	1 x 10-5	11.01	0.018	0.296	2.398	1.018	0.403	4.394	1	265	159	1,157,454
MDD	1 x 10-8	7.46 x 10-8	0.999	1.000	-16.412	18115	-35523	35490	0	265	8	1,157,454
NEUROTICISM	1 x 10-5	2.07	0.475	1.000	0.726	1.015	-1.264	2.715	1	241	697	907,858
NEUROTICISM	1 x 10-8	1.10 x 10-8	0.999	1.000	-18.326	20223	-39656	39619	0	241	84	907,858
SCZ	1 x 10-5	1.42	0.731	1.000	0.349	1.016	-1.641	2.340	1	100	1,042	317,182
SCZ	1 x 10-8	5.41	0.098	1.000	1.689	1.021	-0.312	3.690	1	100	232	317,182
WGL	1 x 10-5	7.82 x 10-5	0.986	1.000	-9.457	521	-1029	1010	0	165	68	891,538
WGL	1 x 10-8	6.51 x 10-7	0.999	1.000	-14.244	22364	-43848	43819	0	165	2	891,538

GWAS Thresh (GWAS significance threshold); OR (Odds ratio); p-value (emprical p-value of the significance of the observed enrichment); cor. p-value (p-value Bonferroni corrected for 16 tests); Beta (effect size from the general linear model, equal to log(OR)); SE (standard error from generalised linear model); CI95_L (lower boundary for the 95% CI of the effect size/beta); CI95_U (upper boundary for the 95% CI of the effect size/beta); No. Annot. Thresh (number of annotated variants within annotation passing GWAS thresh); N.Annot (number of variants within given annotation); N.Thresh (number of variants passing GWAS thresh); No. of SNPs (total number of LD pruned variants). ADHD (Attention deficit hyperactvity disorder); ASD (Autism spectrum disorder); BPD (Bipolar disorder); LOAD (Late onset Alzheimer's disorder); MDD (Major depressive disorder); SCZ (Schizophrenia); WGL (Wearer of glasses or contact lenses).